2013

The hare and the tortoise: the problems with the notion of action in ethics

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

THE HARE AND THE TORTOISE:

THE PROBLEMS WITH THE NOTION OF ACTION IN ETHICS

by

KAROLINA LEWESTAM

B.A., University of Warsaw, 2004

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2013

Approved by

First Reader

C. Allen Speight, Ph.D.
Associate Professor of Philosophy

Second Reader

Amelie Rorty, Ph.D.
Visiting Professor of Philosophy
Tufts University

Third Reader

Paul Katsafanas, Ph.D.
Associate Professor of Philosophy

THE HARE AND THE TORTOISE:

THE PROBLEMS WITH THE NOTION OF ACTION IN ETHICS

KAROLINA LEWESTAM

Boston University Graduate School of Arts and Sciences, 2013

Major Professor: C. Allen Speight, Professor of Philosophy

ABSTRACT

Wittgenstein once asked, "What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?" What would be left is, presumably, the quality of 'agency,' which differentiates between legitimate actions and mere behaviors. In my dissertation I investigate the way we conceive of this quality and recommend a change of the prevalent model for one that is developed in a more empirically informed way.

Most current work in ethics employs a historically acquired and folk-psychology approved notion of agency. On this view, the distinction between actions and behaviors is fairly clear-cut. Actions proper are characteristic of human beings. They are 'rational' in either the deliberative process that preceded it or in terms of their efficacy; they are launched 'autonomously' by the agent's self rather than influenced by context, emotion or habit. These, and a few other conditions have to be fulfilled for an act to earn the badge of an action; falling short of that standard disqualifies it, or, at the very least renders it an imperfect, faulty instance of agency. An agent is thus typically viewed as a disembodied, rational source of conduct, who can withhold her desires and choose between different courses of action using some form of deliberation.

I submit that this model survives neither due to its empirical adequacy nor

because it is otherwise valuable for ethics (or, more generally, for understanding human behavior). Rather, there is (I argue), a certain widespread philosophical *attitude* that determines its persistence—a general longing for the stability of the self and an orderly, controllable relationship between the agent and the world. I call the proponents of this attitude "tortoises" and offer a critique of their main claims. I conclude that we must alter this model. The empirical results from psychology and neuroscience suggest that an agent is best viewed as a bundle of modules that are governed by different rules. None of them is "more" the agent than another, but all operate to achieve a state of homeostasis between so the different processes within the agent and the environment.

ACKNOWLEDGMENTS

I would like to thank my adviser, Professor Allen Speight, for his support and understanding. He is a great person, a fantastic teacher and, on top of that, he is insanely smart. Notice the injustice – some of us do not even have one of those attributes.

I would like to thank Professor Amelie Rorty, whose practical wisdom and sharp intellect can make one believe in Academia again. Things that she wrote inspired me throughout writing this thesis, although I am not sure she would consider this a complement. I pity everyone who did not have a chance to meet her.

I want to give special thanks to Professor Paul Katsafanas, who took the time to help me with this project, even though his life is at the moment incredibly busy. I am very glad I got to meet him—he combines intuitions similar to mine with a superior intellectual skill and a stronger philosophical discipline. He is officially an object of my envy.

Professor David Roochnik, Professor Aaron Garrett and Professor Griswold also deserve my gratitude. In one way or another, they have all contributed to my completion of the program, by accommodating my difficult situation and offering practical solutions. I would like to give all of them a hug, but that would be highly unprofessional.

Nothing would be done, I must say, without the help of my Mom, Aldona, and my mother in law, Łucja. They are sighing in relief as I submit this thesis and promising themselves that they will now try to avoid four year olds for at least six months. I suspect, however, that they will break soon. Orla Richardson helped me beyond belief, too, but she has already fled.

Candice, Irina, Jessica, Vera, Hege, Kirsten, Kenny, Andrea, Miranda—you have been great friends to me here, in Boston, and I love you. Very much. From Cambridge, from Poland, from Tierra del Fuego, from anywhere life takes me.

And Kuba. And Maks.

TABLE OF CONTENTS

INTRODUCTION

* * *

If James is right, and philosophy is to a great extent a matter of temperament,[1] then some divides between us could be more than arguments can handle. The new ruler of Denmark, from Zbigniew Herbert's *Elegy of Fortinbras*,[2] certainly believes so. Looking at Hamlet's dead body, Fortinbras joylessly says, "The rest is not silence but belongs to me," because "I have tasks a sewer project / and a decree on prostitutes and beggars." As Fortinbras goes on with his soliloquy, he keeps flaunting his down-to-earth mentality and pragmatic disposition, fascinated by the contrast between himself and the prince—who was, he triumphantly declares, "not for life". The poem concludes with two haunting lines:

> It is not for us to greet each other or bid farewell we live on archipelagos
> and that water those words what can they do what can they do prince

There are, undoubtedly, many types of temperaments, but in terms of how they manifest themselves in philosophy, that they can be, I suspect, divided into two principal kinds. The first kind, let us name it a Tortoise, likes things slow, yet steady. It seeks

[1] "The history of philosophy is to a great extent that of a certain clash of human temperaments" James, *Pragmatism*, 19

[2] Herbert, *The Collected Poems*. Translation: Czesław Miłosz.

stability and avoids disintegration, takes solace in the presence of a method and shivers at the sight of luck. It prefers rules over habits, and metaphysics over sociology. The lightness of being seems to him quite unbearable, hence the hard, heavy shell on his back. The fear of fracture and decomposition can be, for a Tortoise, best alleviated by strengthening control—be it his control over others, or others' control over him; be it a norm that strongly binds, or indisputable command of God. The self is a unity, repeats a Tortoise; morality is coherent, we have control over ourselves and over our destiny; the outside world probably exists, and there are at least one or two things we can know for sure.

The other kind of temperament is best imagined as a Hare, freer and faster, but without a good *a priori* plan. A Hare has never fully learned to (using Herbert's words again), "Repeat old incantations of humanity fables and legends /…/ repeat great words repeat them stubbornly."[3] Thus when the Tortoise admires a majestic oak tree, or a great philosophical notion, she cannot help but wonder what made that particular acorn sprout, not some other.[4] She is interested in genesis more than the status quo, in the mechanism more than the structure, in the process more than the result. Where the Tortoise sees stability, she sees temporary homeostasis; where he sees nature, she suspects convention. She does not search for design, perfection or unity. She disregards Archimedes' plea that

[3] Herbert, *The Envoy of Mr. Cogito*

[4] This image comes from another one of Zbigniew Herbert's poem, *Dęby*; not yet translated. The poem tell a story of a man who seeks advice from the oak trees, but hesitates to take it, once he remembers their arbitrary beginnings.

a firm place to stand should be given before the Earth can be moved; she grabs at it from a wobbly standpoint and tries to shake it, at least a little bit. The Tortoise is, for the most part, appalled.

The Tortoise and the Hare keep clashing within pretty much all philosophical themes, but the problem of action makes the conflict particularly stark. One the one hand, agency is a candidate Archimedean point from which the entire problem of morality can be lifted, if we take it to be both the metaphysical cause and the normative measurement of the things we do. It is one of the stubbornly repeated great words, a hope for personal unity, a centering notion that, if not mistreated, could allow us to continue speaking in the language of 'the old incantations of humanity'. But, on the other hand, its embroilment in the physical notion of causality invites its own doom from the hands of the Hare, or, if in luck, some half-hearted compatibilistic grace. Neuroscience is almost ready to confirm or falsify the old ideas on how action comes about, and slowly forces metaethics to retreat from the long-occupied descriptive realm into the tiny normative fort. For it seems that increasingly often we are offered new evidence for the lack of control over our conduct, for our inability to compute the right choice of action and for the great influence of body, habit, emotion and context on our daily exercises in agency. Every Tortoise, I believe, already knows that.

In this dissertation I am mainly interested in the accounts of the processes that result in actions. This theme is strongly overlapping with the so-called problem of practical reason, but it goes beyond it—for I am not ready to claim that all legitimate actions must be rational, or that there exists a specific mental process (such as practical reasoning) that

is a *sine qua non* condition for agency. In order to avoid using words that arrive with substantial philosophical baggage, I will simply say that I am interested in the stuff that is going on in the head and its surroundings before (or while) an action happens to happen.

And it happens so, I argue, that most of the mainstream accounts of this 'stuff' are ridden by Tortoise intuitions. This might be not bad in itself, of course; a matter of temperamental preference—but I claim that these inner instincts of a Tortoise temperament often manifest themselves by the tacit allegiance to the habits of language and thought that can be traced back to obsolete frameworks. In particular, I claim that many theories of action are organized as if a certain model of the mind (or a subset thereof, involved in issuing action) were true, namely the computational, von Neumannesque one.

I am not saying that any of the Tortoises I analyze would admit to believing in the adequacy, let alone truthfulness of this model. Most of them would reject it. No one claims nowadays that the mind is best understood as a von Neumann machine, or at least we should hope so. What I mean when I say that these theories have (linguistic and inferential) *habits* that operate *as if* such model was being assumed is a comment on the structure of the metaphorical core of their operative framework. I follow Lakoff and Johnson[5] in their intuition that abstract thought is necessarily metaphorical, and that the mechanics of particular semantic groups can be traced to a system of guiding metaphors, which generate the rules of what constitute a sensible movement of thought in that area. In the conceptual area around the notion of agency, at least for metaethicists, sensible

[5] Lakoff and Johnson, *Philosophy in the Flesh*; Lakoff and Johnson, *Metaphors We Live By*.

movement of thought is organized by the idea that context is input, beliefs are data, and reasoning is symbolic manipulation that generates the answer to the question "what should I do?" On the output we are awaiting an intention, a candidate intention, or, in some specific theories (like Robert Audi's, for instance), action itself.

At the onset of the AI research, symbolic manipulation was considered to be the core and only principle behind intelligent behavior—a paradigm now best known by its goofy acronym 'GOFAI' (Good, Old Fashioned Artificial Intelligence). Not many of the Tortoises have even concerned themselves with the AI research, of course; I am not claiming that this is where they, so to speak, got the idea. This paradigm, however, has emerged from a once commonsensical view on the mind, appropriated for AI by some Tortoise, who was then on duty. It is, in my view, simply the best candidate notion for the 'personification' of the underlying metaphor; an image coherent in itself, and which allows for an effective setup of the discussion between Tortoises and neuroscience, which I am trying to 'arrange'. I am quite aware that the days when rational deliberation in philosophy was openly considered to be a one-to-one function from the available data to the 'best judgment' is a thing of the past; the idea has changed, evolved and became more open to incorporate the emotional, habitual and physical factors in action. But the GOFAI metaphor is still behind many of those evolved theories, visible in what Scheler would call *ordo amoris*[6] of a theory, the structure of the emotional preferences that shimmer through the tendencies of argumentation.

[6] Scheler, 'Ordo Amoris' in: *Selected Philosophical Essays*.

How can this generative template of a secret metaphor be detected? There are no foolproof tools for that. Lakoff and Johnson, in search for their metaphors, simply trace linguistic tendencies, and, very much like ethnomethodologists, probe and poke the semantic taboos, revealing what Richard Rorty called 'final vocabularies,'[7] but in their structural incarnation. The questions we should ask, then, could be these: what does this theorist consider to be the 'problem'? For a Tortoise, the problem would often be the apparent lack of computational efficacy of the currently used deliberative algorithm ("it seems that if we believe that the right thing to do is the *sum* of present reasons, what do we do about this one case, in which two reasons cancel each other out?")[8] Another good question: is her response to the problem *automatically* supplying a missing cog in the current model, or is she even considering switching to a different one? The importance of emotions, for instance, can be a reason for turning to a pragmatic account of conduct, or reinterpreted as value indicators that change the structure of payoffs, and thus influence the result of deliberation. Also, when her framework is 'assaulted' by the realities of cognition or luck, does she speak in a hopeful or a resigned manner? If unpredictability of the world is, for her, not something we take head on, but rather something that we must shield ourselves against with a sufficiently thick theoretical wall, she might be a Tortoise.

And if you *are* a Tortoise, the GOFAI model has a number of added bonuses for you in stock. Even if its taken as a regulative ideal, its very possibility indicates the existence

---

[7] Rorty, *Contingency, Irony, and Solidarity*.

[8] A similar problem can be found in: Lechler, "Do Particularists Have a Coherent Notion of a Reason for Action?".

of rational order. Its in-built logical coherence permits the potential of individual unity and political uniformity, and maintains the hope for the kingdom of ends. It carries within it a ready-made normative mechanism and presumes the world of neatly ordered, 'codable' options. And, when coupled with a certain view on the architecture of the psyche, it provides elegant means to fulfill the need to think about ourselves as unified, while still torn by the contradictory and animalistic forces of the heart and flesh. We simply have to limit 'true' agency to the operations of the device and the one force (separate, or identical with it) that carries out its end result, treating the remainder of motivating forces as a controllable nuisance. And then, even though divided in half, at least we have a good sense of where the 'true' self dwells:

> "…When you deliberate, it is as if there were something over and above all your desires something that is you, and that chooses which one to act on. The idea that you choose among your conflicting desires, rather than just waiting to see which one wins, suggests that you have reasons for or against acting on them. And it is these reasons, rather than the desires themselves, which are expressive of your will…"[9]

The reason I try to show the persistent influence of the von Neumannesque metaphor is my belief that its secret subsistence is harmful for the development of moral philosophy, perhaps even more so than it would be if it was out in the open. It goes, I believe, in the face of pretty much all relevant evidence from psychology, social psychology and neuroscience that we have available at the moment, and nothing indicates that the future will might bring radically different results. The neuroscientific paradigm

[9] Korsgaard, C., *Creating the Kingdom of Ends*, p. 370

of understanding behavior is likely to soon replace the currently reigning Freud-tinted Tortoise perspective as the basis of 'commonsense intuitions' about action. If ethics is not ready to offer a perspective that can be logically, metaphysically and *metaphorically* squared with such view, it might either become irrelevant and ineffectual, or a cause for a cultural neurosis. And it does not need to be this way. When Amelie Rorty in *Social and Political Sources of Akrasia* wrote about the curiously habitual nature of the weakness of the will in modern polity, she could have chosen to simply shame her hypothetical *akratoi* for their failings and mourn their inability to hold on to their principles. This strategy would, however, drive an even bigger wedge between the structure and demands of moral theory and the lives of the 'sinners' she describes. Instead, she chooses to acknowledge the widespread moral schizophrenia as a symptom of the flawed interactions between citizens and institutions, and ponders changes that could foster more desirable, and calmer characters. It is just one small example of how a more adequate perspective on the emergence of conduct allows for a sensible diagnosis of the mismatch between moral demands and culture; the alternative is a pessimistic view of agency and armchair "practical" philosophy.

I begin the first part of this dissertation by introducing the concept of a Tortoise theory of action. I try to show how the fight for more relevant theory should focus on ridding moral philosophy of the underlying calculative metaphor. I then try to show that a Tortoise might try to avoid the naturalistic input because a GOFAI theory of action might be used to generate normativity from what within the paradigm passes as 'description'. I

try to show that this argument cannot be made, and thus sticking to an empirically inadequate model cannot be even justified with philosophical advantages.

The second part is devoted to an exposition of theories that try to model the 'stuff' that happens in our brain before (or during) and action is performed. I focus mainly on Robert Kurzban's modular view of the mind, Daniel Dennett's Multiple Drafts model of conscious brain and Gilles Fauconnier's notion of mental spaces as subunits of just-in-time, non-GOFAI processing.[10] I am trying to use these theories to show a modeling approach that is empirically adequate, explanatory and, as much as possible, continuous from the physiological level to the level of phenomenology.

There are many things that I am *not* going to do. The target of my critique is a general template that happens to shape many conceptions of pre-action processing, and not particular realizations of that template. My argument, therefore, is independent from the debate on what exactly constitutes the data that are being processed (internal or external reasons, calculations of pleasure, moral payoffs from fulfilling different kinds of duties, values of means with respect to ends, etc.) as well as from the particular form of algorithmic manipulation postulated by a theory (be it Kantian universalizability of an adequate maxim, a practical syllogism, the application of moral axioms to problems at hand or a method for weighing reasons). I will therefore stay clear of these debates, with the occasional exception for the discussion of autonomy and its role in agency (my

[10] In the exposition of Fauconnier's conception, I strongly rely on the extended version of the notion of mental space as introduced by Hurley, Dennett, and Adams (Hurley, Dennett, and Adams, *Inside Jokes*.)

interest in the issues of autonomy is prompted by a curious relationship this concept often has with the GOFAI model. The von Neumanesque machine is conceptualized as a processing device, but (for some thinkers) unless the output tape is picked by *someone* (a little homunculus sitting right next to the output slot), and implemented by that *someone* (it cannot simply happen automatically, as the next computational step; an intention has to be formed or the will has to step up, or the self must be clear-headed enough to read off the output tape correctly), there is no agency, only a Chinese room that "speaks" the language of action (with dictionaries replaced by Bayesian tables, of course). The allegiance to the computational metaphor and the need of the control of the procedure often result, I think, in pretty outlandish models of the psyche.

I said I want to argue in favor of a descriptively accurate model of human agency that is built for the purposes of moral philosophy, but still in reasonable agreement with the findings of neuroscience and psychology; a model that avoids the traps of the old language of misguided beliefs about rationality. But this model will not be to a Tortoise's liking: it will view action as a result of a number of different forces, represented by modules that evolved for different purposes. It will see practical deliberation as mostly unconscious, happening across disjointed probes of sparse and unorganized consciousness. It will understand an agent as a Pandemonium of sorts, where different kinds of forces continuously fight for influence, and their winning or losing cannot be well predicted by an algorithm. It will see processing as associative, embodied, habit-driven and parallel, rather than symbolic and linear. And, most of all, it will not need to

use language that postulates some form of homuncular perspective, for as the unifying force it will enlist a form of homeostasis, not coherence, control or the will.

Perhaps, then, if temperaments are the force that divides us the strongest, there is no point in putting Part I and Part II in the same work—for even if a Tortoise accepts the critique and starts to worry about his view, he would not be exactly thrilled by an alternative that violates his deepest moral and aesthetic affinities and his beliefs about the very purpose of metaethics. A Hare might seem faster at first, more suited for the race that the Tortoise, but then, as in the Aesop's tale, she ends up losing anyway, convincing only those that believed her in the first place; the running was useless. We do perhaps, after all, live on archipelagos.

I do, in fact, agree with Fortinbras. There is not much that the water, or the words can do, when *ordines amoris* are radically divergent, and each of us seeks a different kind of solace in philosophical work. And yet, I reckon, the only proper answer to this is, 'Oh, well". We should still try to explain ourselves to the other, for there is no other way of making sense of ourselves as philosophers than the constant rebuilding of a theoretical 'reflective equilibrium' under the gaze of the opponent. Who knows, perhaps Hamlet's problem with agency, his inability to act and the lack of self-understanding might have been alleviated if he had have a chance to speak to Fortinbras, the way Fortinbas speaks to him. We also grow as philosophers by trying to make private intuitions intelligible for differently-minded people (and such is the purpose of writing a thesis). And sometimes, every once in a while, a message in a bottle is found, picked and read. Who knows, maybe somewhere there someone is already building a boat.

At times, however, we are in luck, and we find someone else, who can explain exactly what we have in mind with a skill that, for us, is not quite available. This probably gives our message far more chances of taking some sort of effect. And for that reason I would like to end this introduction with a quote from John Dewey, who happens to expresses the exact concern that has driven me to write a thesis that explores the possibility of empirically adequate modeling of action:

> "It is not pretended that a moral theory based upon realities of human nature and a study of those realities with those of physical science would do away with moral struggle and defeat […] All action is an invasion of the future, of the unknown. Conflict and uncertainty are the ultimate traits. But morals based upon concern with facts and deriving guidance from knowledge of them would at least locate the points of effective endeavor and would focus available resources upon them. It would put an end to the impossible attempt to live in two unrelated worlds. It would destroy fixed distinction between the human and the physical, as well as that between the moral and the industrial and political. A morals based on the study of human nature instead of upon disregard for it would find the facts of man continuous with those of other human beings, and therefore would link ethics with the study of history, sociology, law and economics.[11]

[11] Dewey, *Human Nature and Conduct*, xxvi

Part I

Tortoise in a Warehouse:

Ethics and its Secret Model of the Mind

1. Murdoch's call

Imagine that you are in a shop; a large warehouse, if you will. You have two books with you — one containing your desires, the other one listing your beliefs. You are free, you are clearheaded, you are independent; you are ready to shop. So you pull out your books, consult them carefully (it certainly is great, you think, that your personalized list of beliefs now comes with detailed warehouse inventory included), and engage in the process of calculating all the pros and cons. After a brief period of scratching your head, you come to the inevitable conclusion: given what you know and what you want, you should purchase the reasonably priced juicer from aisle 11. You pull yourself away from the sweets stand, where you were just about to satisfy your cravings, and you head to aisle 11.

A similar image was used by Iris Murdoch as a kind of *reductio ad absurdum* meant to illustrate what has gone awry in the way moral philosophy spoke about action.[12] The frameworks used by thinkers like Stuart Hampshire reduced, Murdoch complained, the psychologically complex process of moral decision-making to an exercise in disembodied, rigidly rule-governed and methodologically uniform exercise in calculation, performed in a world furnished with clearly distinguishable options. In theories of this

[12] This metaphor was originally used by Iris Murdoch in Murdoch, Iris *The idea of Perfection*, in: "Existentialists and Mystics", p. 305. I took the liberty to borrow the idea and extend it. When Murdoch brings up the 'shopping' image, she does it to model Stuart Hampshire's view of moral agency, which she finds to be the problem of philosophical thinking about morality.

sort an agent is presented as a computing device calculating the value of the extensionally presented options: what ought John do if action *x* is a fulfillment of a promise, but action *y* is helping the wounded crash victim, called for by the duty of help? Tell me, John should say, how much is my duty worth against my obligation, and I will calculate the answer for you. But moral agents, claimed Murdoch, do not act in this manner; they are, rather, attempting to remain virtuous against a variety of incoherent drives and cognitive shortcomings, struggling in a mysterious world that, just to complicate things further, bears the imprint of their gaze, a gaze that is laden with habit and emotion. John does not just have two options, but countless options; he could do *x*, but also x_1 or x_2, and more often than not he does not even see some of them as discrete options, blinded by the emotional and cognitive stress of witnessing a car crash. Murdoch thus called for a different strategy in constructing a model for moral decision-making: "A working philosophical psychology is needed which can at least attempt to connect the modern psychological terminology with the terminology concerned with virtue. We need a moral philosophy which can speak significantly of Freud and Marx … (In current philosophy) the will, and the psyche as the object of science, are isolated from each other and from the rest of philosophy."[13]

What Murdoch was trying to say is, I believe, that there is no point in doing metaethics if we start off with an inaccurate understanding of the mechanics of the workings of an agent's mind.[14] The shopping scenario suggests that the processing mind

[13] Murdoch, Iris: *Existentialists and Mystics,* p. 337

[14] I am not saying that Murdoch, or myself, would claim that a thorough understanding of the

is a (distant, but still) cousin of a von Neumann machine, a device designed according to the principles of Good Old-Fashioned Artificial Intelligence (GOFAI[15]) with data and instructions neatly stored, frozen in the expectation of the input question. The difference is that a computing device works as a sum of non-hierarchically organized cooperating parts, while in the mind of an agent there is a supervising element (the self? the will?) that actually pulls all the 'computation strings' and executes the computed outputs. Unless Freud and Marks mess up this all too neat model a little bit, Murdoch believes, its inadequacy will breed more philosophy that will neither be able to guide, nor to explain.

Murdoch identified an important problem with the direction ethics was heading in

---

phenomenon of human mind, on all the levels, is a necessary prerequisite for a responsible theory of action. What we need to have, however, is a fairly good and a fairly accurate model of how we generate action, and that will include, and depend on, some parts of our theory on how the mind works.

[15] GOFAI, an acronym for 'Good Old Fashioned Artificial Intelligence' is a term first introduced by John Haugeland (in Haugeland, *Artificial Intelligence*.) It refers to the idea, prominent at the onset of AI research, that intelligence consists in logical manipulation of symbols. GOFAI proponents would believe that more complex behaviors can be reproduced by the increase in speed and complexity of computing, and by the increase in volume of data, rather than by changes to the architecture of the system. Opponents of this approach would include, for instance, connectionist models of intelligence. A connectionist model allows for non-linear, parallel processing, where more complex behaviors are assumed to emerge from on-the-go feedback-based learning experience which changes the strength of successful connections between 'neurons'.

her time. Whether we are aware of it or not, all talk about moral agency carries assumptions about how the mind works—and these particular assumptions seemed distorting and obsolete. But today things seem different—and Murdoch, I believe, could feel justifiably vindicated.

Since her call, the 'shopping scenario' has earned itself a considerable opposition, which has successfully entered the mainstream debate in ethics. Philosophers have been pointing out the holes in the 'GOFAI plus driver' model, claiming that it is far too simplistic, even when taken as a purposeful idealization. Many of them were of virtue-ethical provenience.[16] They emphasized, for instance, the fact that a shopping model of agential deliberation requires that action choices be uniformly 'codable', identifiable in a discrete manner, as if they were ready-made, uniformly shaped products on shelves (but, "We do not just open doors, leave rooms"—wrote Amelie Rorty—"we leave them ceremoniously, contemptuously, or expectantly. We do not just tend our elderly parents:

[16] I am referring here to the turn towards virtue ethics associated, for instance, with Philippa Foot's work Foot, *Virtues & Vices, & Other Essays in Moral Philosophy*. or Alasdair MacIntyre MacIntyre, *Dependent Rational Animals*; MacIntyre, *After Virtue*.; it is present in Susan Wolf's work Wolf, "Morality and the View from Here"; Wolf, "Moral Saints." as well as Amelie Rorty's Rorty, "Three Myths of Moral Theory"; Rorty, "The Social and Political Sources of Akrasia"; Rorty, "Explaining Emotions"; Rorty, "Moral Complexity, Conflicted Resonance and Virtue"; Rorty, *Mind in Action*.; it also figures prominently in Martha Nussbaum's writings Nussbaum, *Upheavals of Thought*; Nussbaum, *Poetic Justice*; Nussbaum, "Human Functioning and Social Justice," May 1, 1992.

we do so tenderly or exasperatedly, respectfully or resentfully"[17]) They rebelled against the general tendency to represent a good agent as Bayesian rational, lucid, methodical, calculating; and her choices as trivially repeatable due to their calculatively understood rationality ("This idea"—wrote famously Susan Wolf about the allegedly attainable "special state" of pure rationality and impartiality in which, or from which, we are supposedly making moral decisions —"both arises from and perpetuates a false picture of human psychology and value, and it encourages an unduly narrow and ultimately implausible conception of what a correct and rational morality might be"[18]).

Some of the thinkers were put off by the shopping scenario's lack of fit with what we know about the mechanics of human action both from empirical studies and introspection. They demanded that actual human conduct be studied before it is modeled by metaethics. This demand became present even among the scholars associated with the Kantian tradition (David Velleman, for instance, says in defense of his view of practical reason as a methodologically imprecise exercise in narration accompanying the life of an agent: "If this thinking isn't what philosophers call practical reasoning, the problem may be that practical reasoning, as conceived by philosophers, is not something that autonomous agents generally do"[19]). The same concern prompted a return to the Humean vision of the acting agent, which is traditionally empirically informed and based on holistically imagined models of the mind. Jesse Prinz, for instance, convincingly argued

---

[17] Rorty, *Mind in Action*., p. 284

[18] Wolf, "Morality and the View from Here," 204.

[19] Velleman, *How We Get Along*.

that empirical studies indicate that emotions (understood in the Jamesian fashion as 'gut reactions'), not calculative reason, are at the core of moral agents;[20] Patricia Churchland set out to trace the origins of moral decision-making to neurobiology and concluded that what seems to be rationally undertaken exercise in moral agency is often a result of our neurobiological make-up.[21] Since Murdoch wrote her rant against Hampshire, the world around the philosophical agent got more complicated, her self less centered, consciousness more 'gappy and sparse'[22], processing more random and buggy, the mind less susceptible to modeling in a 'GOFAI' fashion, and reason far less mighty and far more connected to the agent's flesh and heart. The shoppers, like sluggish Tortoises, stayed far behind the progress of philosophy, together with those using final causes in scientific explanation or the Cartesian Theatre enthusiasts.

Even those who remained somewhat attached to the image of a free and rational, deliberating shopper at heart, still have refined it and refurbished it, making it more palatable for those concerned with 'Freud and Marx'[23] (though nowadays Murdoch would more likely invoke figures like V. S. Ramachandran, Oliver Sachs or Benjamin

---

[20] Prinz, *The Emotional Construction of Morals*.

[21] Churchland, *Braintrust*.

[22] Dennett's phrase from Dennett, *Consciousness Explained*, 1992.

[23] The refinement of the view of moral agency as connected with calculation of reasons relied mainly on de-universalizing moral problems and moving towards particularism, while retaining the deliberative character of moral reflection Dancy, *Moral Reasons*; Dancy, "Defending Particularism"; Dworkin, "Unprincipled Ethics."

Libet in order to make her point). Ever since Herbert Simon's analysis of the role of context in practical reasoning,[24] microeconomics, the traditional bastion of the Bayesian model, started to admit that not only do agents not deliberate in a manner suggested by the shopping model,[25] but even if they did, they would often be less effective.[26] The shortcomings of the paradigm, it seems, became too glaringly obvious (both in its normative and descriptive aspects) to openly continue doing metaethics in a Wal-Mart setting.

Are we done with the shopping then; can we, philosophers of action, as those that prepare the ground for moral reflection, move toward a more accurate view of moral agency? Can we leave the Tortoise behind and leap forward, towards an adequate explanation of the somewhat messy human action-generation mechanism, which would drop the useless Bayesian idealizations and the habit of seeing GOFAI-like structures as the best idealizations? Are we ready to think about conduct as something that is not a result of conscious calculations, but that organically emerges from the phylogenetic constraints of our biology, the context of our moves *and* our application of culturally

---

[24] Simon, *Administrative Behavior, 4th Edition*., see also Simon et al., *Economics, Bounded Rationality and the Cognitive Revolution*..

[25] This view gained most respect with the empirical findings of Kahnemann and Tversky (see for instance, Kahneman, *Thinking, Fast and Slow*.)

[26] This particular thesis is most recently championed by the 'abc' research group from Max Planck institute in Berlin (see, for instance, Bouissac, "Bounded Rationality"; Gigerenzer, *Adaptive Thinking*.)

developed conceptual tools? I claim that this is not quite the case yet. Scratch the surface of the prevailing discourse in ethics today, and an image similar to Murdoch's will show.[27] There will be a computational device with a driver inside, a mind that calculates options like a Bayesian machine, with payoffs specified by moral rules—and there will be a little homunculus on top of it, in charge of the process and execution, taking care that all this fuss is happening *for the benefit of someone*, or is done is done *by* someone.

My goal in this part is to re-expose the shopping view as the persistent metaphorical understructure of our thinking about practical reflection. I want to look at the way we sort and utilize common (and not so common) intuitions in mainstream metaethics, and show that the current direction projects a primitive model that ultimately devalues the achievements of moral theory.

I do not argue, however, that this 'GOFAI plus driver' model is openly endorsed. It is rather enduring only as a deep and unconscious structure that governs our theories of practical thinking. Much like the cognitive theorists Lakoff and Johnson (who are often quoted, but hardly taken seriously) , I believe that most thinking is to some degree metaphorical, and that, specifically, "It is virtually impossible to think or talk about the mind in any serious way without conceptualizing it metaphorically."[28] They identify a number of GOFAI-related structures that we apply to thinking about the mind, such as 'A

[27] This is, curiously, less true about meta-ethics, and more true about ethics, as if the theory of how we should conceive of agency had little effect on the actual conception of agency employed in ethical inquiry.

[28] Lakoff and Johnson, *Philosophy in the Flesh*, 235.

Line of Thought is a Path" (linear view of processing) or 'Ideas are Entities with Independent Existence" and 'Ideas are Locations' (beliefs are discrete data, stored in long-term memory).[29] Lakoff and Johnson believe that the ultimate source of metaphors that are central to our thinking is the bodily experience[30]—and that is, to simplify their view, why for upright, fall-averse bipeds like us 'up' will always be better than 'down' across pretty much any sort of semantic field. There is a deep and obvious truth in it. Perhaps the persistence of the GOFAI model which I profess is not a sign of a misguided allegiance, but a testimony to some deep fact about our biological nature. This possibility certainly needs to be investigated—and, in case it were true, the benefits of retaining the computational model should be reexamined, and its status renegotiated.

But I also believe that Richard Rorty was right, to an extent at least, when he criticized philosophers for their all-too-desperate (and yet mostly unconscious) allegiance to "…literalizations of what once were accidentally produced metaphors,"[31] suggesting that vocabularies can change, thereby changing the central system of metaphorical templates that users of language depend on in the structuration of their though. At least some of the metaphors can be gradually replaced, or rendered less potent, provided that there is a good reason for that (in Rorty's system the reason would be somewhat Hegelian—a moment might come where a particular vocabulary is no longer useful, fitting or inspiring, and a poet, or a philosopher, has an opportunity to reimagine the most

---

[29] Ibid., 236.

[30] Lakoff and Johnson, *Metaphors We Live By*.

[31] Rorty, *Contingency, Irony, and Solidarity*, 61.

basic metaphorical structures). Even if there is a justification from biological determinism for this particular metaphor, we might still be able, even if only to a certain extent, to assist Rortian 'poets' such as Freud and Marx, in finishing their quest for a non-computational vision of the practical mind. We should actively try, in fact: if a paradigm is being openly rejected, obviously empirically inadequate, explanatorily unhelpful; if we also have the belief that there is a value in empirically informed ethical models related, among other things, to the guiding potential of a theory, we have all the signs we need to conclude that that the residue of its influence should be exposed, tackled and, if possible, replaced.

2. Tortoise (and Hare, too)

Before I go on to tracing the problematic model in existing theories of practical thinking, I want to spend some time on putting a slightly more vivid face on the enemy.

Let me call those who are relying on the GOFAI model of practical reason 'Tortoises'. The famous Aesop's tale, quoted at the beginning of part I, presents the Tortoise as the good guy: he sets out to win the race, and, due to the unrelenting nature of his commitment to the game, ends up first at the finish line. The Tortoise is a von Neumann machine, with linear processing of action-related data—not the most efficient way to figure out what to do, but a sure way to get you where you want to be (eventually). The premise of his practical reasoning, namely that he needs to win the race, remains a stable reference throughout the way—there are, it seems, no parallel processes happening in his little reptile mind. Slow, not very relatable, and yet boringly reliable, the

Tortoise embodies the GOFAI view of decision-making mechanism that produces successful action.

Contrary to popular opinion, Tortoises are not very easy to spot. A Tortoise, even when pressed, will likely not confess his true commitments to a certain vision of the agential mind. Thus it might be helpful to gather some clues as to what overt claims might possibly indicate that we have just met a Tortoise.

Let me go back to Murdoch's image of a shopper as a helpful tool to identifying a Tortoise view. The shopping scenario conveniently contained three things: first, some assumptions about agent's psychology (she was acting 'freely', she was able to withhold action until her deliberation was done, her beliefs and desires belonged in different books, etc.) Second, it hinted at the existence of a method of determining the right course of action (weighing pros and cons), and, third, it (metaphorically) presented the context of action in the form of an organized warehouse. Though likely interdependent within a theory, these three kinds of claims have separate foci. Thus at times, it will be useful to think of them as separate; for those occasions let me label them t-psychology, t-method and t-landscape, respectively. These labels need not to be rigidly defined; they are just broad labels that serve as a pragmatic tool for pointing at some aspects of particular theories.

Tortoises are likely, then, to make certain kinds of claims about human psychology—these will be their assumptions about the way the human mind functions when going about its agential affairs. These claims will have descriptive character and are most obviously vulnerable to arguments from experimental psychology and neuroscience.

Sometimes they will be made overtly, for instance when M. Bratman begins his analysis by saying that we are, as humans, capable of making plans[32]—that is a straightforward descriptive claim about the way we are wired. But they will also often be carried within a normative assertion: if a hypothetical Tortoise, for instance, submits that "A good agent performs Handel's *Hallelujah* before each action," she probably remembers G. E. Moore's observation that "ought" implies "can,"[33] and thus she effectively claims that one *can* in fact sing *Hallelujah* when an action calls. This, depending on how exactly she sees a satisfying performance of the oratorio, can imply that agents are capable of memorizing tunes, are generally not tone-deaf, or that their motivating drives can be successfully suspended and not acted upon for the duration of the singing.

A Tortoise knows, like everyone else, that humans do not always seem cool-headed, rational or cognitively capable of performing the GOFAI type of deliberative exercise. Thus the claims of t-psychology will typically be aimed at squaring the shopping image with the impurities of actual conduct. Since our minds at first glance seem to be not only capable of deliberation, but also of silly humor and imagining unicorns, a Tortoise will likely claim that practical reflection is effectively limited to a small subset of cognitive affairs. Jokes and unicorn-related musings will be called the lesser denizens of the thinking machinery, while the von Neumannesque element will be glorified, encouraged, and (often) elevated to the position of the 'true' agent.

Since a Tortoise encourages the usage of the said subset as the vehicle for action-

[32] Bratman, "Reflection, Planning, and Temporally Extended Agency."

[33] G E Moore, *Moore Ethics*., see also: Russell, "Ought Implies Can."

related thinking, he must also assume that such move is psychologically possible. One of the ways in which we can spot a Tortoise is by his distrustful attitude towards those elements that are often effective in conduct, but are hard to incorporate in a computational framework. Traditionally, these elements were passions, habits, automatic actions, etc., basically those elements that can move us to do things before the computing process is done. Thus a Tortoise must favor the idea that temporary suspension of motivational forces is possible—or, that there can be room made for computational deliberation. That suggests, since desires will be conceived as relevant information for the computational effort, that they can be conceived as movement-independent informational bits, which can be used for the calculation of the right course of action. A Tortoise will generally speak of both, desires and beliefs, as independent units of that sort, which can be retrieved from memory in order to feed the processing device. In other words, Tortoises will often allude that mental contents are propositional in form, and functionally divided (into conative and cognitive), neatly stored in separate cabinets.

"T-method" will refer (again, roughly) to this subset of the Tortoise view that is concerned with the procedure that must be used for making the right choice of action—the deliberative method. Since the procedure, or the algorithm, is at the very center of the GOFAI model, it will be, one can assume, pretty important for a Tortoise to give that procedure a prominent place in action modeling. Hence Tortoise frequent preference for procedural view on rational action: they can easily believe that agency can be secured by following a certain method, an algorithm or a set of steps. When Kant asserts that free agency can only be achieved if the will adheres to a universalizable maxim, it is not the

mere feature of that maxim's universalizability that makes it right for an agent to act on it. It is rather the agent's deliberative journey that must expose this fact about the maxim in question; only once deliberation is properly performed, the will can be genuinely motivated by the so discovered duty.[34] The von Neumann machine must be, then, set in motion, used, rather than handed the right answer.

Deontological universalizability is perhaps the most famous example of a procedure that, if correctly followed, produces genuine agency as the output, but most other theorists have a version of such a procedure, too; they range from the Bayesian calculation, through practical syllogisms to 'weighing reasons'. A Tortoise is, however, not stupid, and knows that sometimes even the most carefully designed procedures seem to produce undesirable results. Here, in order to detect a Tortoise, we must look for the general attitude towards such occurrence. Is the candidate Tortoise bemoaning this fact, lamenting, Calvin-style, the wretched design of the fallen creature, who is too flawed to use the light of reason properly, because "The light still shines in the darkness, but the darkness comprehends it not" [John1:5]? Is he looking for ways to render the algorithm more precise by increasing its complexity (such as inclusion of context-indexication, encoding emotional processes, etc)? Either reaction indicates that we might have just met a Tortoise.

Now, if we think of a Tortoise agent as an avatar in a computer game of adventure, t-psychology will describe its pre-established capacities, such as the ability to stop before the next step is decided. For t-method, a player could look into a

[34] Kant, *Groundwork*

‘walkthrough’, where helpful tips specify how the next courses of action should be determined (for instance, “killing a werewolf is worth 50 points, while killing a dragon is worth 100. You have only 20% chances of killing a dragon, while there is 60% chance you will successfully fight the werewolf. Focus on werewolves until you collect the golden hatchet and increase your chances with the dragon by 40%”). In this analogy, “landscape” will be all that furnishes the game’s fantasy world: the way the missions are structured, the rules that govern the physics, the way in which some fragments of the world are salient (if werewolves can kill you, they are salient; if mockingbirds cannot be interacted with and fly the digital skies for solely aesthetic purposes, they are probably not salient). Each philosophical game of rational agency comes with a landscape of sorts: actions are either metaphysically discrete or their borders can be gerrymandered through a choice of description. The game might require you to be in motion lest you will be killed, or it provides an opportunity to stop, save and think before the next move is due. Your results can be a direct function from certain variables, such as correct application of a procedure or a number of minutes devoted to reflection, but there also can be luck programmed into the scheme in various amounts, and all you can do is trying to diminish the chances of being eaten by werewolves. The salient elements might be limited to the agent herself, but they can include other agents, social norms, physical forces. T-landscape is, then, the tacit model of the world on the basis of which a Tortoise thinker built her view of practical reflection.

A Tortoise, since he is committed to the GOFAI metaphor of generating action will likely see the world as furnished with discrete options—only in this way the best

course of action can be computed. He will speak in a language that presumes, or alludes to the existence of fixed values that mark each of the choices. Metaphysical realism about values is one way to secure getting the properly formatted data for the von Neumann machine, but in fact anything that provides fixed values will do—for instance, as internalists would prefer, a relation of a given course of action to the agent's motivational set.

Here is an important structural fact: it is immediately visible that the claims within each set are not necessarily logically connected, and neither are particular sets of claims with one another. As a Tortoise, you can be picky and treat them as distinct theoretical units: you can, for instance, pick some claims from t-method and renounce t-psychology altogether.[35] Neither set is a theory in the logical sense of the word (it would be, if it contained n claims *and* all their logical consequences). They are, one can say, torn bags filled with pretty random stuff, tied together with the general direction of reference. And, as a Tortoise, you can engage particular claims within the sets in different combinations and to different degrees, mostly without risking logical contradiction. That said, the modules still are commonly (through philosophical habit, for instance) co-dependent, frequently mutually reinforcing, and often, in concrete Tortoise theories, *presented* as logically connected.[36]

---

[35] As rational action theory in economics does, for instance.

[36] It is not difficult to see how that happens. The belief in the existence of specific psychological faculties or properties might easily determine the way you view deliberation—but on the flip-side, sometimes it is your faith in the efficacy, or rightness, of a particular way of decision-

Does it not seem, then, that in a situation like this it would be better to deal with sub-Tortoises, classes of theories that subscribe to logically cohesive sub-parts of the view? Even though I assert that most of the Tortoise claims listed above stem from the underlying reliance on GOFAI type of modeling, the danger is that it is too vague of a connection. And if one is not convinced by my suggestion that these types of claims are connected by the way they fit together into a metaphor, my critique might resemble a very silly war on all that were born in July, have an uncle in the military and a particularly itchy scalp.

I cannot fully escape this objection; all I can do is to put all my cards on the table and admit that (i) there is no way of 'proving' my intuition that a metaphorical commitment unifies all these claims in some way; a certain amount of evidence is all there can be provided,[37] and (ii) one of the reasons for my belief in the plausibility of the label is a feeling I have about the psychological (not just structural) coherence of these kinds of theoretical commitments. A quote from James explains the sort of feeling I rely on, when (later in the chapter) I put together, in one category, such different thinkers like

---

making that prompts you to assume a specific view of mind's moral apparatus. Any theory of proper practical reasoning contains a tacit model of reality—we have to, at the very least, have a way to individuate between choices of action in order for the concept of deliberation to make any sense at all.

[37] Lakoff and Johnson justify their findings regarding the central metaphors behind abstract concepts by the analysis of large quantities of text; I will try to provide samples of philosophical claims later in the chapter.

Korsgaard and Perry, or Bratman and Berker:

> The history of philosophy is to a great extent that of a certain clash of human temperaments. […]Temperament is no conventionally recognized reason, so [a professional philosopher] urges impersonal reasons only for his conclusions. Yet his temperament really gives him a stronger bias than any of his more strictly objective premises. It loads the evidence for him one way or the other, making for a more sentimental or a more hard-hearted view of the universe […] He *trusts* his temperament. Wanting a universe that suits it, he believes in any representation of the universe that does suit it. He feels men of opposite temper to be out of key with the world's character, and in his heart considers them incompetent and 'not in it,' in the philosophic business, even though they may far excel him in dialectical ability.[38]

I believe that James is right, and that temperaments are the main force behind the construction of a philosophical theory. And I also think he is right to say that it is the *clash* of temperaments that generates conflicts more strongly. One of the methodologically unacceptable, and yet undeniably convenient tools that I use to detect the common thread among those that I named 'Tortoises' is 'emotional probing': I put myself against the candidate theories, and the unity of the feeling of a sudden, deeply felt emotional clash, which I experience reading them. This, of course, has no theoretical value, and those, who are not convinced, have a full right to reject the idea that the label 'Tortoise' is useful. Perhaps they will become more sympathetic as we go on. If accepted, however, this assumption makes my project less of a silly war on a randomly gathered

[38] William James, *Pragmatism*, World Publishing Co., 1970 (original 1907). P. 19

people and more like a study on individuals with abdominal pains, weekly panic attacks and low self-esteem, where logically independent properties turn out to be connected genealogically (resulting, in this particular case, from clinical neurosis).

I am, thus, convinced that many seemingly independent commitments of a Tortoise can be construed as a consequence of a motivational set characteristic of a certain type of philosophical attitude, or temperament. This assumption has methodological consequences, too—it justifies the search for extra-theoretical motivations that could explain seemingly independent commitments.[39] But then another question arises—who are we really after? Those with a reptile temperament, or theories that use GOFAI as the underlying model of processing that leads to a choice of action? My belief is that the two things are strongly tied together: a philosophical attitude of a certain sort will, I argue, gravitate towards frameworks that 'sit well' with the GOFAI model. The promise of a unifying and communicable method for choosing actions, and the unity of the agent suggested by the linear character of the GOFAI processing, together with the normative nature of rule-governed reasoning that this framework implies, cater to those temperaments, who fear chaos and uncertainty. The *ordo amoris* of a Tortoise places rules over phronesis, uniformity over personality and answers over questions. Their fear is the disintegration of the self, and their concern is self-unification;

[39] One might ask now, why not define a Tortoise with these 'extra-theoretical motivations', but through the claims that are only their byproducts? The answer is that it would be disingenuous: I can only see claims, and from them I can speculate about motivations, but they are not the element that is given.

their love is stability and their worse nightmare is moral schizophrenia. Max Scheler wrote, "Whether I am investigating the innermost essence of an individual, a historical era, a family, a people, a nation or any other sociohistorical group, I will know and understand it most profoundly when I have discerned the system of its concrete value assessments and value preference […]*Whoever has the ordo amoris of a man has the man himself* […] he has a spiritual model of the primary source which secretly nourishes everything emanating from this man."[40] A Tortoise's *ordo amoris* is not well hidden. She is constantly nourished by the vision of a stable, controllable method of action choice, incorporated into the very nature of the acting self, the kind of practical reason that could have been modeled on the old-fashioned AI conception of intelligence. When she must give it up, she gives it up with sadness, and whenever she can bring it into a theory, she does so with joy.

I have been defining a Tortoise depending on her claims *and* the motivations for accepting them, hoping that my exposition will be a sufficiently good guide for phronetic classification. But at times I might be useful to know that there could be a sharper tool of classification. For those occasions, I will assume that all of the components, t-psychology, t-method and t-landscape, consist of a perhaps large, but nonetheless finite number of claims. Let us call the set of all these claims T_0. Note that T_0 is a fairly wild set, where all minor and larger claims are accepted, including claims C_1 and C_2, where C_2 is a logical consequence of C_1. Imagine that we now can use an algorithm that would purge T_0 of all the unnecessary claims like C_2 and end up with a more concise set of

[40] Scheler, "Ordo Amoris" in: *Selected Philosophical Essays*, 100.

claims T.[41] Now we can conceive of the term in question as a kind of cluster concept, where a theory is Tortoise if and only if it contains no less than a certain number of these claims (I, of course, do not know the exact number, but I can say this: it is definitely at least three and we might determine it with time. And I can also say that more often than not every Tortoise will accept claims from at least two modules, but that is *not* specified by this definition, especially because the borders of modules are vague and some claims, once rephrased slightly, can float). [42]

---

[41] We could use the following algorithm: we pick a claim (for instance C_n) from T_0 and put it in T. Now, we pick another claim from T_0 (C_m) and check whether C_m is a consequence of C_n. If yes, then we do not take it in, and we move to the next claim from T_0. If not, we check whether C_n is a consequence of C_m. If yes, we exchange C_n for C_m, if not, both are in. For each next tested claim, both steps have to be repeated for each claim already in T. And, of course, we go on until we attended to each claim in T_0. One serious advantage of defining T through this algorithm is that is takes the burden of unbearable neatness in submitting candidate Tortoise claims — I am only building T_0.

[42] Note that this definition abstracts from the motivational considerations, so wildly advertised before as that what lies at the heart of a Tortoise view. This should alert the reader to a possible problem with my analysis (I, of course, consider it merely a *feature*). If all Tortoises can be effectively picked out solely by the claims they make, *and* a Tortoise has certain set of motivations for holding these claims, one of the following must be true: (i) in a strike of extensional felicity, it is a great (and pure) coincidence that everyone accepting these claims also happens to have certain motivations that I am describing, (ii) these claims necessarily cannot be held without that kind of motivation, (iii) I am making a fairly strong assumption that these

And now a plot twist: this particular disjunctive definition comes with one necessary condition, a *sine qua non* for admission into the Tortoise club: every Tortoise, as I will explain later, believes that there are genuine 'actions' and 'non-actions'. The definition can, thus, be sufficiently well expressed in the following way (with Θ denoting the property of being a Tortoise theory, and C is the property of containing a specific claim (1, 2, 3, etc.) from set T):

$$\Theta(Th) \Leftrightarrow C_1(Th) \wedge \{[C_2(Th) \wedge C_3(Th) \wedge C_4(Th)] \vee [C_2(Th) \wedge C_3(Th) \wedge C_5(Th)] \vee [\ldots.]\}$$ [43]

That will be, I hope, sufficient for understanding the idea of 'Tortoiseness'. But what I would most want the reader to keep in mind as we go on is the guiding image of a shopper. The image, that is, of the disembodied shopper, easily replaceable with any other similarly rational one; a shopper with externalizable, propositionally conceived

---

particular claims are mostly (but not necessarily) held because of such motivations. (i) is obviously barely possible, and even if it were, no advantage for understanding a Tortoise would come from thinking about her motives; (ii) is very improbable, so it must be (iii). I am making the assumption that certain kinds of philosophers, given the historical and psychological contingencies that brought all of us here, are very likely to hold on to certain claims *because* of certain motives. And I *am* guilty, I am sure of them, of misattributing and overgeneralizing this motivations. More often that not, I am imagining a more coherent, Tortoise, a Platonic version, if you will, and re-imagine his character.

[43] The number of disjuncts is finite and it lists all three-element, four-element (all the way to n-1 element, where n is the number of claims in T) combinations from T.

cognitive contents, who suspends her needs and wants for the time of deliberation, and whose choice, though procedurally conducted, is an inevitable destiny—a function from beliefs, desires and the one good method. It offers a touchstone of what a Tortoise will like and dislike about a theory of agency, and what she will try to preserve even in radical improvements. Incomplete as this image is, it gives us (I hope) an imperfect tool of understanding a Tortoise's theoretical moves, the meaning of her trade-offs and the direction of her hopes, and thus equips us with both, the boldness to challenge her deepest assumptions and the respect for her impulses.

3. The secret shopping

The task at hand is, then, to find traces of Tortoises in metaethical thought, and show that they (i) exist, (ii) can be plausibly described in terms of reliance on the GOFAI metaphor, (iii) argue against the GOFAI view of the practical mind. How can one go about finding them? Following Lakoff and Johnson, and their method of extracting the guiding metaphors from language, I should probably begin by taking an inconspicuous piece of philosophical text that concerns moral choice, and see what is implied by the way it is written.

Consider, for instance, the following exposition of a morally charged situation by philosophers Kearns and Star:

> "Imagine you are walking along a busy road, deep in thought, and someone suddenly pushes you over… Suppose this person tells you that he has just saved your life by pushing you out of the way of a fast-traveling bus ... He also tells you that he almost didn't push you and thereby save your life because he is in a great

> hurry to get to a meeting. You thank him profusely, and he hurriedly goes on his way. It seems that a *natural* [emphasis mine] way of describing what just happened in this scenario is to say that this stranger considered the reasons for and against helping you and he made a quick judgment that, given the balance of reasons in play, he really ought to save you. No description of this kind would be appropriate if the person who pushed you over had simply accidentally tripped.[44]

It is hardly a surprising way to begin an article. This kind of framing of moral problems seems to have remained the same since Hampshire was told off by Murdoch. What can be unpacked from this quote?

First, the authors' emphasis on the thinking process implies that the natural state of an agent is a Tortoise-like stillness of decision making, not practical engagement with actual action. So many moral dilemmas of philosophical protagonists play out in slow motion, where a drowning baby can hold its breath long enough to survive through the agent's efforts to weigh the pros and cons of jumping into the lake. It is a sign that processing is considered as more salient than, and separate from, performing. Second, the processing of perceptions, if we take this way of speaking at its face value, is utterly divorced from the situation at hand and considers stable data sets (like, for instance, beliefs about respective moral values of the competing action choices). The final decision emerges at the output slot as the result of the calculative process that took place in an abstract von Neumann space, where there is no time pressure or emotional stress, and where cognitive overload is only a scary story told to sleepy neurons at the end of the day. Third, actions are described as discrete options, tokens of types, and thus

[44] Kearns and Star, "Reasons," 31.

underdescribed—we do not know, for instance, whether the stranger rescues you "exasperatedly, respectfully or resentfully."[45]

But perhaps it is too harsh to accuse theorists like Kearns and Star, who after all have a reasonable particularist bent, of a tendency towards a GOFAI-type of decision-making model. It is after all nothing like, say, Audi's firm perspective on practical deliberation:

> The overall theory of practical reasoning proposed takes practical reasoning as an inferential process with both motivational and cognitive premises. It corresponds to a practical argument, which, in turn, is a kind of argument appropriately produced in answering a practical question. Practical reasoning is indeed an inferential realization of such an argument"[46]

Perhaps, but even particularists, Murdoch's natural allies, who are supposed to be sensitive to the organic and hardly formalizable cooperation of the agent and her context for the issuing of action, are not free of the secret Tortoise love for the computational approach. Or maybe even not *that* secret. Below, Andrea Lechrer defends particularist conception of reasons against Selim Berker's accusations of the lack of coherence. Berker is, according to Lechrer, guilty of assuming that there can only be *one* combinatorial function (which "…takes as input the valence and weight of all the reasons present in a given possible situation and gives as output the rightness or wrongness of each action available in that situation"[47]) for the calculation of the net reason for action in

---

[45] Rorty, *Mind in Action*., p. 284

[46] Audi, *Practical Reasoning*, 188.

[47] Berker, "Particular Reasons," 103.

a given situation. But, apparently, there are

> "…two functions. One of them (what he calls the "total reason function") yields for each action the total reason in favor of the action on the basis of the relevant contributory reasons. The second function determines whether an action is right or wrong on the basis of the total reason in its favor and the total reason in favor of alternative options. For instance, one's total reason function might tell one to add up the weights of all the reasons in favor of an action and subtract from this sum the weights of the reasons against the action. The second function, in turn, might specify that the right action is that with the highest amount of total reason in its favor."[48]

A particularist, on this view, does pretty much the same thing that the kind stranger in the quote above, except her task is a little harder—the possible actions are still extensionally described and with values attached, but their final worth depends on complicated relations with a number of context-specific variables.

The bizarre thing is that this sort of computation seems arduous and ineffectual: it appears to be both, computationally taxing *and* hardly useful for increasing the moral, or other, worth of the subsequent action—let alone phenomenologically familiar. Were the brain asked to perform such exercise every time a choice is called for, the energy expenditure would be dangerously high. The redundant character of adequately diligent 'weighing' reminds me of a character named Dodecahedron, from Norton Juster's book for children *Phantom Tollbooth.*[49] It is a story of a boy named Milo, and a very useful piece of literature, because it offers the kind of absurdity that does not derive its power

[48] Lechler, "Do Particularists Have a Coherent Notion of a Reason for Action?".

[49] Juster, *The Phantom Tollbooth*.

from playing with chaos,[50] but from highlighting, augmenting and utilizing the tacit structures and metaphors that habitually govern our ways of thinking. Dodecahedron suddenly appears when Milo and his friends are trying to choose one of the three roads leading to the city. When Milo asks, "Perhaps you can help us decide which road to take", he officiously delivers, in the form of a question:

> "By all means (…) There's nothing to it. If a small car carrying three people at thirty miles an hour for ten minutes along a road five miles long at 11:35 in the morning starts at the same time as three people who have been travelling in a little automobile at twenty miles an hour for fifteen minutes on another road exactly twice as long as one half the distance of the other, while a dog, a bug, and a boy travel an equal distance in the same time or the same distance in an equal time along a third road in mid-October, then which one arrives first and which is the best way to go?"[51]

Now, that is a long way to say which way is the longest, especially when they are, it seems, all the same in length. I have a nagging feeling that by the time the net reason for action is determined or Dodecahedron is done, anyone in their right mind would already be driving down one of the three roads for just about *any* available reason (and by the time Kearns and Star's sweet stranger weighs all his considerations, your wife is in the bank, cashing your life insurance).

Fine, one might say; perhaps there is something about the ubiquity of your Tortoises, and the resulting persistence of the computational view of the mind in ethics. But here are two good objections to your interpretation of the meaning of this fact. First, it might be

---

[50] Lewis Carroll's *Alice's adventures in Wonderland* would be an instance of that, I guess.

[51] Juster, *The Phantom Tollbooth*, 127.

more peripheral than you make it to be. You yourself said that the GOFAI model has been updated and developed; instead of taking the easy way of focusing on hardly representative quotes, you could look at one of those accounts that truly attempt to reject the old fashioned calculation; if you found the GOFAI way there, it would be at least somewhat informative. Second, you might be accusing some theories of descriptive faults, when they are in an entirely different business. Even if ethics *does* sometimes speak of agents as if they were calculating devices that process the values of the extensionally described courses of action in the light of the stored data such as beliefs and desires, they might be engaged in one of the following practices: (i) building an idealized model of action-generation mechanism for explanatory purposes, (ii) building a distorted model with the valuable feature of cross-subjective communicability, (iii) constructing, not describing (not even in the sense of structurally adequate model-building) an essentially political tool that facilitates beneficial uniformization of behavior-related institutions.

Let me start with the first objection. I want to argue that the shopping image is what generates and maintains the theoretical development of most mainstream theories of action. But yes, there are notable instances of theorists that go against it, and try to introduce the right measure of psychology into a computational perspective on the action-generating apparatus.

Consider, for instance, Nomy Arpaly—a philosopher who built her career on providing thoughtful arguments *against* the standard GOFAI model, a true hero of the anti-Tortoise movement. If we believe in the efficacy of the computational mind, we

should generally desire that an agent follows whatever the system judges to be the best thing to do, all things considered—or for her to follow the 'best judgment'. There has been a lot of philosophy written on the best judgment issue—and the driving force behind the concern with the best judgment must be the idea that there is a fairly precise function from the set of beliefs and desires of an agent to an appropriate action choice in a given context.

The best judgment is a clever device: it relativizes the rightness of the choice of action to the particular agent and her epistemic situation (what is not in the data storage cannot possibly be 'considered'), while in the same time providing the taste of universality of the method—a prerequisite for any self-respecting GOFAI model (should you have the same choice, same beliefs and goals, you must make the very same choice, insofar as you are rational). Arpaly observes, however, that the procedure for arriving at the best judgment is not always getting you where you want to be (say, in aisle 11 of our shop, right by the product that is *so* right for you). Oftentimes, she claims, it appears that the best thing to do is to act *against* one's best judgment; that a 'fit' of irrationality can be a better choice than complying with the results of careful deliberation.

That is a hopeful nod towards a non-computational view of the agential mind, a recommendation to drop the Dodecathedron's way and choose the right road with your gut. The anecdotal agents starring in Arpaly's argument are people, who formulate the best judgment according to the rules of the *techne*, but end up acting otherwise. Her description of a hypothetical agent Emily goes in the following way:

> Emily's best judgment has always told her that she should pursue a PhD in chemistry. But as she proceeds through the graduate program, she starts feeling restless, sad and ill-motivated to stick to her studies. These feelings are triggered by a variety of factors, which let us suppose, are good reasons for her, given her desires and beliefs, not to be in the program. The kind of research that she is expected to do, for example, [would] not allow her to fully exercise her talents, she does not possess some of the talents the program requires, and the people who seem most happy in the program are very different from her in their general preferences and character. All these factors she notices and registers, but they are also something that she ignores when she deliberates about the rightness of her choice of vocation: like most of us, she tends to find it hard, even threatening, to take a leave of a long-held conviction and to admit to herself the evidence against it. But every day […] her restlessness grows, her sense of dissatisfaction grows, and she finds it harder to motivate herself to study. Still when she deliberates, she concludes that her feelings are senseless and groundless. One day, on an impulse, propelled exclusively by her feelings, she quits the program, calling herself lazy and irrational, but also experiencing an inexplicable sense of relief. Years later, happily working elsewhere, she suddenly sees the reasons for her bad feelings of old, cites them as the reasons for her quitting and regards as irrationality not her quitting, but, rather, the fact that she held on to her conviction that the program was right for her as long as she did.[52]

Arpaly does seems to respond to Murdoch's call for a moral philosophy "…which can speak significantly of Freud and Marx." She is remarkably respectful of Emily's (and, by extension, any agent's) natural psychological mechanisms, and, instead of squeezing Emily into a all too tight and energy-expensive model, she extends it so that this clearly-not-insane Emily can be comfortably and duly accommodated. Thus one can choose to see Arpaly's effort to veer away from t-method as a sensible adjustment of a theory that neglected some features of human psychology, to wit, the difficulty in

[52] Arpaly, "On Acting Rationally Against One's Best Judgment," 504.

accessing relevant data from within the mind. A von Neumann machine is self-transparent, and so is our shopping friend from the beginning of this chapter—but Arpaly's Emily is not. It seems we have arrived at an example of a theory that does not see computational processing as the natural skill of the mind, and, by extension, as the regulative ideal for deliberation about action.

There is, however, a slightly different, and no less plausible perspective on what is going on. First, Arpaly, like so many other theorists of rational action,[53] clearly *appreciates* Emily's struggle to formulate the best judgment. Her concern is not that Emily approaches the problem in the wrong way; rather, she is worried that the method of Bayesian calculation, while the best we have got, is not fool-proof when used by a psychologically flawed creature. No matter how many times Emily tries to formulate the best judgment, diligently following a prescribed step-by-step process, there will always be cases where the structure of her psyche will render the whole process quite useless, and the outcome will be unsatisfying. Why is it the case? Emily, like all of us, is a victim of Freud: her mind sometimes refuses to release important information. The connection between her long-term memory and working memory is sometimes clogged, and for every ten easily retrievable beliefs, there will be one or two firmly jammed in its log-term

[53] Kahnemann's view comes to mind here as an example of a theory that points at the incongruence between the GOFAI view of deliberate action and the unfortunate shortcomings of *actual* agential behavior. In Kahnemann's world, every struggle towards Bayesian rationality is valuable; every achievement of the computationally correct choice is an occasion to rejoice. Kahneman, *Thinking, Fast and Slow*.

memory drawer.

Were Arpaly not influenced by the old myth of GOFAI processing, she would have ended her story here, saying that there is no algorithmic method for choosing action rationally, full stop. She might have added that the concern with an action's rationality will not take us very far in terms of theory of agency, or even in terms of adequate self-assessment and the judgment of others. Perhaps, to illustrate the unfortunate but not eliminable vagueness of agential guidelines she would have had quoted Freud himself, who turns the problem with the inaccessible data into a theory of the mind, which then served as a justification for an intuitive 'method' of decision-making:

> "When making a decision of a minor importance, I have always found it advantageous to consider all the pros and cons. In vital matters, however, such as the choice or a mate or a profession, the decision should come from the unconscious, from somewhere within ourselves. In the important decisions of our life we should be governed, I think, by the deep inner needs of our nature"[54]

Instead, Arpaly is so disappointed with the sudden loss of a function that infallibly determines the right action from the set of desires and beliefs that she takes her GOFAI processing elsewhere: outside of an agent. She will remain a Tortoise at least partially, with t-psychology renounced, but with t-method intact.

> "There are at least two ways to think of theorizing about rationality. One way is to see the idea theory of rationality as providing us with a manual of sorts: follow these instructions, and you will always make a rational decision […] Another way is to see

[54] Freud, cited after Dijksterhuis*, Think different: The merits of unconscious thought in preference development and decision making*.

> theorizing about rationality as aiming […] at providing us simply with a theory — a theory that tells us when people act rationally and when they do not, so that given a God's eye view of a person's circumstances, beliefs and motives, one would be able to tell how rational or irrational said person would be in performing a certain action. These two fascinating tasks—to which I will refer as the creation of a *rational agent's manual* and the creation of an *account of rationality*—are more different than they look."[55]

Arpaly's dissatisfaction with Emily's well-reasoned choice (it simply does not square well with what Arpaly believes is the right thing to do) leads her to abandon the quest for the 'manual' as a lost cause, and switch to the 'account of rationality', where the problems of psychology are no longer relevant. The two tasks, I believe, are not as different as Arpaly would want us to think, at least they are not that different in the way she tackles them herself. The 'account' is but a vision of an idealized agent, who happens to be fully lucid and absolutely self-transparent, with the ability to process the accessible data according to the moral algorithm—and in the case of this sort of agent the 'best judgment' method (described in the 'manual') would indeed be fool-proof. We are still in the same shop, with the same two notebooks, except that this time they have coffee spilled all over and we are perhaps better off, for now, ceding the shopping onto an assistant, who kept his copies of our data intact. Let him be the agent we all want.

In Arpaly's account, actions are, discretely, laid out in front of an agent as possible choices. There exists a method of arriving at the right choice that can be conceptualized in an algorithmic manner and stored in the computational device like instructions in a computer. And the GOFAI model has been taken out of the mind and

[55] Arpaly, "On Acting Rationally Against One's Best Judgment," 488.

posited as a regulative ideal, towards which the flawed agential machine should strive.

It seems, therefore, that the recognition of some psychological features of agents does not safeguard a theorist from the commitment to a GOFAI-inspired view of agential mind. Arpaly exported the von Neumann machine outside of Emily and her lot, and hand it to God, or a bird and her view, as a measurement of their performances. But she could also have a less aspirational view on how action should be generated, and do what Jesus would—embrace human processing device with all its flaws. Perhaps to model practical intelligence *truly* means to model Emily, not an abstracted, self-transparent entity. The problem is that it would be easily doable. GOFAI architecture of data processing can be quite accommodating. To simulate Emily on a von Neumannesque machine we can make just a few little adjustments: introduce an instruction that sometimes shuts the processor's access to a small subset of data and, in cases when the decision at hand requires some input from that closed subset, change a method of computing the best judgment from a Bayesian one to random, or vaguely probabilistic.

This is precisely why so many 'improvements' over the shopping scenario are in fact cases of replicating the old fashioned AI paradigm in modeling—we can always imagine a Bayesian interpretation of a new input or structural limitations. We can, for instance, recognize the positive, or necessary role that emotions play in decision making—but instead of changing the deep structure of the decision model to include emotional forces, we can treat them as a challenge to our coding method. We could, for instance, translate them into a function that systematically adds value to certain outcomes.

It seems, then, that Tortoise sympathies can endure through even quite radical framework adjustments. But *cui bono*? Who benefits from maintaining the primary metaphorical pull at the core of the language of rational action? What part of the Tortoise psychology is hung up on the linear, computational perspective on practical thinking?

My guess is that a part of the temperamental make-up of a Tortoise is the desire for controlled politics of agency; a desire for a standard that will simultaneously safeguard one from moral luck, and put adequate pressure on the citizens of the Tortoise land. A GOFAI framework, with its computational demands and one-to-one function from the set of beliefs and desires to the best judgment about action provides a Tortoise with a tool that secures the firmness of the rules and allows for their imposition. When we demand that agents deliberate according to an algorithmic standard, even vaguely specified, we gain a normative tool to rule the T-kingdom, simply because, "When practical reasoning meets the minimal standards of adequacy, the reason it provides is normative—and so a reason *for* one to act—by virtue of the content of its concluding judgment."[56]

An alternative to the firmness of some kind of computation is, I reckon, a somewhat disgusting prospect for a Tortoise. Aurel Kolnai in his beautiful essay *On Disgust* suggested that moral disgust detects not wrongness, not blameworthiness, but amorphousness, an internal incapability of strong acceptance of certain hierarchy of

[56] Barker, "Audi's Theory of Practical Reasoning," 44. I am not saying (yet) that it is a bad standard, or an unjustified one; I am simply wondering what makes us want a standard of this sort.

values, lack of firmness, insolidity of the soul.[57] A Tortoise, conceivably, cannot give up the idea about an algorithmic key to moral agency entirely, without a theoretical nausea—especially when the alternatives seem to entail the dissolution of a moral agent and her moral community.

That is, however, pure speculation. The purported authoritarian longing of the Tortoise mindset might, nonetheless, become slightly more pronounced when we analyze the Tortoise need to distinguish between *true* actions and *non-actions*. A stringent theory of agency, as I will try to show in the next section, can become a way to create a somewhat plausible psychology, where the von Neumann machine does not seem like a strange implant, a *deus ex machina* in the otherwise unruly theatre of the mind, but a *machina ex deo*, a device fully integrated into the arrangement of the soul.

4. Actions and Shmactions:[58] generating normativity from the notion of action

Many people believe that, as far as things go, some things are actions and some other are not. A Tortoise is among them—for her the result of the Wittgensteinian subtraction ("What is left over if I subtract the fact that my arm goes up from the fact that I raise my

[57] Kolnai, *On Disgust*. On that view, moral disgust is a true cousin of physical repulsion, which also detects putrefaction, decomposition, insolidity, misplacement and lack of structure in organic matter.

[58] This funny term is borrowed from David Enoch (Enoch, "Agency, Shmagency"), and means an action that is only *seemingly* an action, but in fact does not stand up to the standard.

arm?") is always a positive number. It is, I claimed before, a *sine qua non* condition of joining the Tortoise squad.

In this section I will try to show that the Tortoise approach to this distinction is *not* motivated by a genuine inquiry into the mechanics of action, but aimed at the preservation, and justification, of the kind of normativity that springs from the commitment to the GOFAI model. This section is thus aimed to motivationally deconstruct the alleged 'descriptive' component in Tortoise models, thus further justifying the need for exchanging it to a more empirically adequate one.

Is the belief that there are actions and non-actions in any way problematic? At first glance, at least, it makes more than perfect sense. Most theorists of action, in fact, both from the side of Tortoises and from the hare party, share this intuition. Consider, for instance, the uncontroversial way in which Donald Davidson begins to think about agency. "This morning," he writes, "I was awakened by the sound of someone practicing violin. I dozed a bit, then got up, washed, shaved, dressed, and went downstairs, turning off a light in the hall, as I passed". Further on, he pours himself some coffee, stumbles on the dining room rug, and then spills some of the coffee while reading the *New York Times*. "Some of these items record things I did; others, things that befell me," he wonders, and adds that in the attempt to sort these things into the two categories, "Many examples can be settled out of hand, and this encourages the hope that there is an interesting principle at work […]"[59]

"Yes," we might agree—"It does make sense!" But even though Davidson's

[59] Davidson, *Essays on Actions and Events*.

words read so smoothly, one thing, I reckon, should not be missed in this quote: there is, he says, a '*hope'* for 'a *principle'*. The seeming regularity of, and intuitive confidence in linguistic behavior is enough for Davidson to trigger the immediate faith in a deeper integrity of the notion of action. If this hope is fulfilled, Davidson's getting up and turning off a light will not be 'actions' for different reasons, as (we can easily imagine) they could be—they will turn out to be actions because *they share a set of features*.

This is, of course, an inclination toward order that we all have; something that keeps us going as theorists. If Davidson was a scientist, we could say that sensitivity to regularity is a desirable part of the context of discovery, something that makes discovery possible at all. But what if this hope—the *essentialist hope*—becomes so strong that the 'sensitivity' is replaced with headstrong conviction that *a principle will be found*? It is, I believe, the case with Tortoises.

One can argue that even if those bad Tortoises indeed do not respond to linguistic regularity with adequate caution, it might perhaps warrant a correcting effort, rather than a general methodological worry. Assume that the conditions for actions they will offer will be too stringent, too broad or somewhat off target. Still their pursuit of the 'principle' could be similarly laudable as that of Linnaeus, whose categorizations of species was enormously helpful for the development of Biology, even though he was unaware of the essential fuzziness of the very concept of species—something that we are acutely aware of in a neo-Darwinian world. There are (perhaps only in a sense) actions and non-actions, just like there are (again, in a sense) fish and jellyfish—why would it be wrong to point that out (or try explain how to adequately sort them)? Who has a problem with that?

I actually might, provided that someone goes beyond neutral classification of this kind, and adds that jellyfish are, by definition, superior, and thus my worth is inversely proportional to the number of my scales. I argue that, for a Tortoise, an evaluative opportunity opening up with the kind of definitory exercise she employs seems to be a strongly motivating reason to give in to the essentialist hope. Tortoises are Tortoises *qua* metaethicists, not *qua* folk-psychologists, neuroscientists or law practitioners. They are, therefore, directly involved in the normative inquiry. They are not asking, what *they* can do to understand human behavior, they are asking what a theory of action can do for their main concern. They are not in this business because they are passionate about human action generation mechanisms and the secrets of raising arms; they are hoping to buy the shares of the enterprise out and invest it in the already pre-ordered concepts of the right and the good. It is as if Linnaeus wrote his *Flora Lapponica* with the sole purpose of showing mosses that they are inferior to lichens, and started off by defining mosses as "those things clearly superior to lichens". In a similar manner, a Tortoise hand-pick those shoppers who suit her purpose and uses them to postulate an essence that will be used to shame the non-complying shoppers.

How is this done? To find an example we do not have to venture far. I will focus on the strategy of argument employed by Christine Korsgaard, one of the most prominent Tortoises writing today. Korsgaard, writes, for instance: "Action is self constitution," and adds, without a blink of an eye, "…accordingly, what makes actions good or bad is how well they constitute you."[60]

[60] Korsgaard, „Self-constitution: Agency, Identity and Integrity"

The interesting thing is that Korsgaard starts off with what seems to be a normatively neutral definition of action, perhaps an analytically derived one, perhaps a result of generalization from observations[61]—and, in a movement presented as logically inevitable, ends up with a norm. Here is another instance of this strategy:

> "Why is disunity a threat? Why is unity essential to agency? Unity is essential to agency, whether collective or individual, because an action, unlike other events whose causes in some way run through an agent, is *supposed* to be a movement, or the effecting of a change, that is backed by the agent as a whole."[62]

Once again, what action is *supposed* to be (a unifying activity) defines the possible paths of moral failings (doing something that threatens the unity of an agent). In *Self-Constitution in the Ethics of Plato and Kant* we find yet another instance of the very same argument:

---

[61] Korsgaard never explains whether she considers her definition to be analytic or synthetic *a posteriori*. At times she writes as if her conception of action was a generalization from particular instances of intuitive appraisal of acts (or at least these instances serve here as justifications)—and hence the analogy to Linnaeus. Her Kantian affiliation could, however, suggest that it arises from purely conceptual analysis. My critique is more relevant if the former is the case; fortunately, even if she intends to do the latter, the unintuitive result of such analysis keeps her from putting her cards on the table, and forces her to seek confirmation from particular representative 'cases' (Korsgaard, "Self-constitution: Agency, Identity and Integrity", "Self-Constitution in the Ethics of Plato and Kant.")

[62] Korsgaard, *The Normative Constitution of Agency*, p. 8. Emphasis mine.

> The Constitutional Model implies a certain view about what an action is, which in turn has implications about what makes an action good or bad. […] [it] tells us that what makes an action yours in this way is that it springs from and is in accordance with your constitution. But it also provides a standard for good action, a standard which tells us which actions are most truly a person's own, and therefore which actions are most truly *actions*. [63]

A standard of a behaviors fit with an agent's overall constitution is, then, again generated from the mere appraisal of the phenomenon of agency. In a similar vein, in Korsgaard's interpretation the categorical imperative is binding because it falls out of the idea of action ("The categorical imperative is an internal standard for actions, because conformity to it is *constitutive* of an exercise of the will, of an action of a person as opposed to an action of something within him"[64]).

It is by no means a new way of obtaining a normative component from the analysis of (metaphysical or linguistic) facts. It places us back in the familiar Aristotelian world, where only once we knew who we were, we were able to know where to go.[65] The

---

[63] Korsgaard, "Self-Constitution in the Ethics of Plato and Kant," 3.

[64] Ibid., 27.

[65] Korsgaard has recently been admitting to Aristotelian influences affecting her interpretation of Kant's ethics (Korsgaard, *Self-Constitution*; Korsgaard, *The Constitution of Agency*; Korsgaard, „The Origin of the Good and Our Animal Nature"; Korsgaard, "The normative constitution of agency"). She found, it seems, a way to use some of Aristotle's ideas to slightly naturalize Kantian ethics, making it more palatable for wider audiences and common sense (specifically, she suggests that ethics as conceived by Kant is a fitting conceptualization of the kind of functioning that is specific to humans and makes them "do well" in life, and thus it squares with "Aristotle's

idea, as used by Korsgaard, can be summarized in the following way: in a slightly modified, post-Aristotelian vernacular[66] we might say that, at least in certain cases, the knowledge of what X is provides us with a set of conditions that we can use to evaluate how good is a certain y such that X(y). In other words, if being a knife means to be able to cut, a blunt knife should be sharpened—and to get that "should" we did not have to venture anywhere beyond the knife itself. What a relief, since we are, as meta-theorists, on the level where normativity should be found, not appealed to. Korsgaard, in David Enoch's words, believes that "…the normative standards relevant for actions will fall out of an understanding of what is constitutive of action, just as the normative standards relevant for cars fall out of an understanding of what is constitutive of cars."[67]

If this belief were true, venturing into theory of action would be a pretty good bargain for a Tortoise. Korsgaard seems to have successfully outmaneuver both, a skeptic about morality and Humean methodological police (the public force that keeps 'ought' and 'is' apart). "What do you mean there are no moral standards?" says Korsgaard to the former, "Surely you agree that a knife is better sharp (and, while we are on it, action is better as self-constitution)?" To the latter, she might say, "I committed no crime, because

---

idea […] that the good for a being consists in the well-functioning of that being as the kind of being that it is, in circumstances that are conducive or favorable to its overall well-functioning.", *The Origin of the good and Our Animal Nature*, p. 1)

[66] For the most famous instance of the adoption of this idea, see for instance: McIntyre, *After Virtue*—but also all the Korsgaard references listed above.

[67] Enoch, "Agency, Shmagency," 170.

my 'ought' is just another expression of my 'is'. I looked at action, and I saw it for what it was; now I have an uncontroversial tool to measure your agential performance. As easy as judging knives."

Surely, a knife is better sharp. But if (counterfactually) most of knives were in fact sharp, this analysis would have meager normative power. A sharp knife in the land of sharp knives could only shrug dismissively if a moralist instructed it "Be sharp!". We have a justifiably strong sense that Korsgaard is first, a seeker of normativity, and only second an action theorist, and thus this is, for a Tortoise like her, a difficult problem. She cannot give in to a definition of action that endows the term with too broad a denotation. That is why she must reject what she calls, "The naturalistic conception of agency – that an agent is active when her movements are caused or causally guided by her own mental states or representations."[68] That category includes, for instance, Davidson's theory.[69] If she was to agree with him that action is any event that under some description is 'intentional,' than she would either need to preach sharpness to a legion of laser-sharpened chef knives, or she would have to give up the entire plan of obtaining a norm from the essence of action. Same goes for, say, equating action with consciousness of one's movement,[70] and many other conceptions formed through an inquiry that is not norm-sensitive. "Agency is not just a particular form of causality"—explains her dissatisfaction Korsgaard in a somewhat circular manner, "because causes, just as such,

---

[68] Korsgaard, *The Normative Constitution of Agency*

[69] Davidson, *Essays on Actions and Events*. Essay 3.

[70] Searle, *Intentionality*.

cannot succeed or fail. It is not immediately obvious how this feature of the concept of agency can be captured by an account that explains agency in terms of the causality or causal guidance of an appropriate mental state."[71]

A normativity-seeking Tortoise, therefore, is not going to succeed by merely adopting the controversial maneuver of abandoning genuine description in favor of questionable teleology. The stratagem can only work if action is defined in a way that puts agents under adequate pressure. That is why, for a Tortoise, action must be far divorced from non action. Her definition will, thus, always be *aspirational*: by no means a description of something we *do*, but something that we should do—preferably something that squares well with a Tortoise's prior normative intuitions about what it means to morally "succeed or fail". It makes sense, says Korsgaard, "For when we think of agency as something that is normatively constituted, *the very idea of an action has a certain honorific character*"[72]

There are two problems with this set-up. The first one is the fairly obvious circularity of this strategy, initially advertised as uncontroversial in terms of generating normativity. Despite the Aristotle-inspired attempts at naturalization of normativity, Korsgaard assumes the standard that she sets out to justify with a definition as a part of that definition. The second, larger problem is that of theoretical identity. I imagine that despite the "honorific" and aspirational view of agency, a Tortoise metaethicist like Korsgaard does not think of herself as someone coming to work as an attorney of a

[71] Korsgaard, *The Normative Constitution of Agency*, 10

[72] Korsgaard, *The Normative Constitution of Agency*, 4. Emphasis mine.

standard to be justified at all costs. She must believe that her reflections on the nature of agency are genuine and disinterested; she must believe that her description is referring to an authentic phenomenon. At the very least she would view an attack on her view that pointed out its inability to account for day-to-day agency as legitimate. But then again, talking about raising arms and taking showers rather than acts of duty cannot take her normativity quest very far—and she is back to the aspirational view.

Rarely is this problem is addressed head-on.[73] Usually, it is solved by means that are largely rhetorical. The principles of Tortoise rhetoric instruct them to keep arm-raising in mind, but talk mostly about being *extremely sharp, incredibly keen*, so the norm will remain to appear as tied to some sort of description. The Tortoise action-related literature is, therefore, focusing mostly on 'real' actions, 'actions proper' and 'true expressions of agency', 'self-governing agency' and 'genuinely autonomous actions'—seemingly descriptive regulative ideals compared to which 'normal' actions can now be found wanting (but explained *qua* actions, the way Plato's shadows are explained by their resemblance to an *eidos*).

Consider, for instance, the work of Michael Bratman. He habitually uses this sort of rhetoric, trying to concurrently maintain the intuitive advantages of calling arm-raising an action and the normative convenience of a strictly aspirational definition. He does talk about daily conduct, but that, to him, is not strictly speaking something of philosophical

[73] In Korsgaard, "Self-Constitution in the Ethics of Plato and Kant" Korsgaard is trying to get out of the normatively charged definition of action conundrum by saying that "bad actions" are, in a way, actions too; they are "The same activity, badly done" [p. 15]

interest. He insists instead that humans are capable of *more* purposeful actions, *real* actions (such as, for instance, self-governance) and writes, "A central problem in the philosophy of action is how to understand these *stronger* forms of agency."[74] He frequently refers to a certain sub-class of actions as 'full-blown': "…when action is the issue of normative deliberation anchored in such an endorsed conception of practical identity that there is full-blown agency and not merely an outcome of causal pushes and pulls."[75] Here, in a fascinating rhetorical move, Tortoise Bratman reduces the universe of activity to 'super-actions' on the one hand and 'causal pulls' on another—because too much attention to the middle of the continuum would hinder the exploitation of action theory's normative potential, available only if in the mind of the reader 'action' and 'super-action' get somehow confused. Richard Foley used the same trick, when he wrote, "Actions that are done not only intentionally but also deliberately can be regarded as actions *par excellence*, […] [they are] paradigms of actions."[76] Sarah Buss talks about "Super-actions"[77] This way of speaking allows one to simultaneously feel that they are in fact trying to understand all actions for what they are and reject causal theories that are, in fact, doing just that.[78] And, with all these maneuvers, Wittgenstein's exercise in subtraction is seamlessly exchanged for a different puzzle: what should I *add* to me

---

[74] Bratman, "Three Forms of Agential Commitment," 329. Emp. Mine.

[75] Bratman, "Two Problems About Human Agency," 317.

[76] Foley, "Deliberate Action," 54.

[77] Buss, "Autonomous Action," July 1, 2012.

[78] I do not mean to declare my support for causal theories here.

raising my arm to make it a *real* action?

We have, then, the conviction that actions are distinct from non-actions;[79] this, in turn raises the hope for constitutive normativity for agents—which, consequently, leads Tortoises to formulating 'aspirational' conceptions, skewed toward normatively charged analysis. My goal was only to expose the motivation behind a Tortoise's drive to first, clearly divide actions and non-actions, and secondly, to cling to this folk-psychological term as a ring buoy for their normative needs that only works when explicated as a badge of honor rather than a property of behavior.

It is worth mentioning (although it is ultimately a little off topic) that the post-Aristotelian call of alleged metaphysical destiny has, in fact, a meager normative force. The most succinct criticism of this strategy comes from Enoch:

> "At times Korsgaard writes as if she believes that the threat that your inner (and outer) states will fail to deserve folk-theoretical names (such as "action") is indeed a threat that will strike terror into the hearts of the wicked. But no support is offered for this surprising claim. And notice that Korsgaard's problem here is not merely that the skeptic is unlikely to be convinced by such a maneuver. The problem runs deeper than that because the skeptic should not be convinced. However strong or weak the reasons that apply to him and require that he be moral, surely they do not become stronger when he realizes that unless he complies with morality his bodily movements will not be adequately described as actions."[80]

Enoch is right, I think, that the skeptic, when designing a house to his liking,

---

[79] The definition of action can function either as a clear-cut categorization device, or as a touchstone for the "degree" of agency present in a particular behavior.

[80] Enoch, "Agency, Shmagency," 108.

should not be thrown off his course simply because someone says: "This is *not* what a house truly is!" With a normative admonition of this sort the problem is not really about defining a house, but giving me reasons to actually build a house rather than whatever I set out to make. In Enoch's words, "…if a constitutive-aim or constitutive-motives theory is going to work for agency, then, it is not sufficient to show that some aims or motives or capacities are constitutive of agency. Rather, it is also necessary to show that the "game" of agency is one we have reason to play"[81]

For a short moment let us imagine, however, that the constitutive stratagem could work—that the principles of metaphysics have the power to confine moral destiny. Our personal metaphysician, a Tortoise, defined action in a way that makes it quite hard to be an agent (she needed that as a warrant for normative demands). What happens when the bar for 'fully-blown' and 'very real' agency is set too high, not just for an ordinary Joe, but for pretty much all the Joes out there?

In short, nothing. Once again we can see how the Tortoise strategy sacrifices philosophical goods, such as empirical accuracy, genuine explanation of existing phenomena and the possibility of actual guidance of real agents, for normativity. When

---

[81] Ibid., 186. I think the response from Aristotle-tinted Korsgaard would be that of teleologically determined inevitability of the search for happiness, which can only be fulfilled if the right way of life (that of self-constitution and duty) is followed. I think we can safely say that it is, in fact, a question the answer to which is at least in part empirical. Evolutionary psychology together with experimental psychology are probably in a better position to determine the solution of a problem so posed.

Kant wrote, "In actual fact it is absolutely impossible for experience to establish with complete certainty a single case in which the maxim of an action … has rested solely on moral grounds and on the thought of one's duty,"[82] he essentially confessed that this worry, for a Tortoise like him, is irrelevant, and if people cannot be proper agents, that their problem, not the theory's. In her *Skepticism about Practical Reason*, Korsgaard argues the Kantian point about the coextensiveness of moral agency and rationality, and similarly concludes that, "The extent to which people are actually moved by rational considerations, either in their conduct or in their credence, is beyond the purview of philosophy. Philosophy can at most tell us what it would be *like* to be rational."[83]

The puzzling move of grounding the normative in the description of "action" while renouncing the relevance of the description of what agents actually *do* when they act is not only a failed theoretical strategy, but also an expression of a curious metaethical attitude. How stringent, and how removed from the daily reality our norms are is not simply a matter of taste. Moral philosophy does not exist in a vacuum; a theorist cannot responsibly provide practical norms without taking an interest in the political, or psychological consequences of following them. If it is true, that moral philosophy is, essentially, either shaming or guiding. Doing the latter rather than the former depends on accepting the nature of the guided subject. This is not done, apparently, by a nod to Aristotle's view of morality as a part of human nature, at least not when "human nature" becomes all too Kantian. It is not done by deriving morality from philosophy of agency—

[82] Kant, *Kant*.

[83] Korsgaard, "Skepticism About Practical Reason," 25. Emphasis mine.

at least not if philosophy of agency is done by someone who already is a moralist. A Tortoise, thus, is immune to the sentiment expressed best by John Dewey, who claimed "What cannot be understood, cannot be managed intelligently."[84]

When in your *ordo amoris* struggle is valued higher than mastery, and tough love higher than understanding, it is yet another sign that you might be a Tortoise. It is by no means an irrational view. From a certain perspective, there is something strangely comforting in seeing ourselves as stretched between the right and the mundane. Oliver Letwin describes this attitude as "philosophical romanticism":

> "The hallmark of philosophical romanticism is the belief that the human condition is permanently and irredeemably unsatisfactory (...) Like his artistic counterpart, [the philosophical romantic] takes the view that we can never be completely at home in the world because our "true selves" are, in one way or another, compromised by the circumstances of our existence. He insists that our life involves ultimate disjunctions — between what we are and what we wish to be, between our feelings and our reason, between one aspect of our existence and another"[85]

So far I looked at the strategy of "constitutive normativity" — defining the category of action, and using this definition as a normative touchstone for candidate behaviors. There is another strategy of that sort: instead of focusing on actions, a tortoise can focus on the 'bearer' or action, the human being, and from the species-specific features derive the normative claims of agency. Fortunately for a Tortoise it seems that we are the only ones with a fully functional von Neumann device in our head, with the

---

[84] Dewey, *Human Nature and Conduct,* pp. xx

[85] Letwin, Oliver (1987). *Ethics, Emotion and the Unity of the Self*. Croom Helm, New York.

little 'self' crouching at the output slot. At first this strategy might not look much different from the strategies described above. After all, with her construal of Aristotle, Korsgaard did assume that agency is constitutive for being a human; her idea of what an action is determined both, the understanding of our purpose as creatures and our morality. But what was woven into the organic structure of Korsgaard's more general view can be considered separately, as a distinct argument—and one that is far more widespread in the philosophical universe.

The problem of the relevance of our biological category for understanding agency is strongly connected to the problem of the division between human beings and the rest of the animal kingdom, and thus tends to stir up a lot of emotion. I will, therefore, illustrate this argument in a somewhat unorthodox manner: through the exposition of a strongly analogous argument from aesthetics. This will allow me to attend mainly to the very structure of the argument, leaving the associated philosophical embroilment aside.

The argument in question is old and simple. Actions, a tortoise ponders, are done by humans. Perhaps, then, the analysis of action could be a natural byproduct of the analysis of the property of being a human. And then—she hopes—we could perhaps harvest practical norms not from the essence of *action*, but from the essence of the *medium* that brings action about. Whatever is human-specific, it would be the very thing that should be present when we *truly* act (of course it would be beyond great if we were the only ones with the capacity for rational processing, but let us pretend that this is not what is being assumed just yet.)

A similar tactic was widely used in aesthetics under the name of "medium

essentialism;" I will try to summarize it briefly, following the lead of Noel Carroll.[86] The problem is not that of agential mastery, but that of artistic value. Instead of asking how we can adequately tell actions from non-actions, aestheticians ask: how can we adequately assess to what extent and whether at all an object is 'art'?

Some theorists of cinema[87] would say that artistic value is strongly dependent on the degree to which a given object exploits medium-specific features; the more film-like a film is, therefore, the better it is aesthetically. Carroll describes his years as a graduate student of film theory as the time when this way of thinking was never questioned:

> "It seemed self-evident at the time that the best films were the most cinematic, that they were the best because they were cinematic, and that if anything were to succeed as a film, it would be necessary for it to employ the peculiar features of the so-called medium"[88]

Certain 'cinematic' directors, like Hitchcock, were considered better then those that were not (like Bergman). "Sometimes we called these other directors 'literary,' says Carroll, "It was not a polite way of speaking."[89] Art is, on this view, a predicate earned through obeying the nature of a medium. If we assume, for instance, that the essence of film is movement, a film made entirely of stills (like Chris Marker's *La Jetee*[90]) should

---

[86] Carroll, *Theorizing the Moving Image*, Carroll, "The Power of Movies."

[87] Such as Rudolph Arheim and André Bazin, for instance. I am only referring to the debate about film here; medium essentialism argument, however, can be applied to any artistic medium.

[88] Carroll, *Theorizing the Moving Image*. P. 1

[89] Ibid., 1.

[90] This is an exaggeration, however —"La Jetee" incudes at least two seconds of moving images.

impolitely be called "photographic", and its value as art would be questionable. Similarly, the fact that humans might be capable of specific kinds of conduct might justify the judging of conduct with respect to the presence of these specific features. This view has been first articulated by Thomas Aquinas, for whom the distinctively human capacity was the capacity to reason. Since for Aquinas agency was only present in *voluntary* acts,[91] he had no problem concluding: "An act is voluntary when it is an operation of reason,"[92] thus effectively narrowing down the scope of action to acts committed by humans. Reasoning that begins with the biologically defined 'medium' and ends with what makes conduct valuable is, since then, omnipresent. Here is a good example of that, just to come back to my favorite tortoise:

> "Moral standards are standards that govern action, so if it is true that *only* human actions have a moral character, and it is true that *only* human actions are governed by reason, it seems plausible that these two properties should be associated. That is, there must be something quite distinctive about human action – something that makes our actions different from those actions of the other animals – which in turn explains why human actions, and those alone, can be both rational and subject to moral governance."[93]

Now, medium essentialism is composed of two distinct theses. Just like in the constitutive normativity strategy, we have a descriptive claim (for instance, "Unlike other

---

[91] See, for instance: Pink, "Reason and Agency."

[92] Summa Theologiae (1265-74) 1a 2ae q6 al

[93] Korsgaard, "That short but imperious word ought: Human Nature and the Right". Emphasis mine.

media, film allows for two-dimensional moving image") and a normative claim that is supposed to follow from it ("Film should exploit movement"). If it worked, we would have a great tool for measuring artistic value of films, and a sense of what an ideal film would look like (in my extremely simplistic example it would be enough to attach a camera to a racing car). "I see you are a film" — I imagine an aesthetician interrogating *La Jetee* — "You must know that, unlike photography or painting, you are capable of actually *moving* your images. Oh, you are *not* doing that? What do you mean you have other merits?? These are things that photography could do.

Similarly, if *unlike animals*, we, humans, have the capacity to (insert a faculty, a skill or a mental state *x*), *therefore* we should strive to (use a faculty, hone a skill, induce a mental state *x*), acting "like a beast" would be an ultimate failure. In an interview, Korsgaard once said: "…Everyone who reflects must ultimately come to see her humanity itself as an essential and foundational feature of her practical identity." [94] Thus, when she begins her explanation of Kant's psychology by saying:

> "A non-human animal is conscious […] But we are self-conscious, which means that we represent to ourselves not only things in our environment, but our own mental states: we think about them"[95]

… it comes as no surprise that self-awareness and meta-cognition turn out to be essential for agency. Meta cognition, of course, performed as a computation over an increased set of data, which now include also symbolic codes of our first-order

[94] Korsgaard, interview

[95] Korsgaard, "Motivation, Metaphysics, and the Value of the Self," October 1, 1998, 50.

cognitions.[96] "You want to be a good agent" — the interviewer inquires this time around. "And yet you have rescued your child from a burning building without even asking yourself: shall I do this? Did you even turn on the device in your head? While there might be some value in what you have done, keep in mind that a dog could save her pups in this manner".

Christian natural law theorists and Kantians are by no means the only ones for whom the description 'distinctly human' is tied with 'normatively required'. There is a part of this argument that seems to be self-evident for nearly everyone involved in thinking about action. Consider for instance Harry Frankfurt's initial opposition to the metaethical version of 'medium essentialism' as philosophically unsound. Frankfurt begins his explanation of the difference between agency and behavior by contrasting a freely walking spider with a different spider, whose legs are moved by strings attached to them by a mean boy. "Wait, so the first spider is an agent, just like us?"—medium essentialist would ask in disdain. Anticipating this, Frankfurt writes:

> "But we must be careful that the ways in which we construe agency and define its nature do not conceal a parochial bias, which causes us to neglect the extent to

[96] Here is a good place to explain why I stuck with personifying the computational metaphor as a von Neumann machine, not a Turing machine as such, or a Harvard one. Once we demand that deliberation i salso *of rules*, or happens on a 'meta-level', we are, essentially, computing both data and instructions. A von Neumann machine can do it, the other ones cannot—it is the only model that enables the storage of data *and* instructions together.

> which the concept of human action is no more than a special case of another concept whose range is much wider"[97]

Perhaps, hypothesizes Frankfurt, action is *not* what remains after subtracting lifting one's paw from lifting another one's arm; perhaps there is something about both of these acts that makes them either an action or a behavior. He thus subscribes to a version of causal account of agency, normatively neutral and abhorred by Korsgaard. So far, so good, a hare might say—but this very un-tortoise move leaves Frankfurt without the possibility of an uncontroversial normative assessment. Thus he fixes it promptly developing a view of personhood, which ends up being not much different from Bratman's view on "full-blown agency". On this view, the admission into the circle of creatures that can perform morally relevant actions requires that we are capable of having higher-order conative attitudes towards our own desires. It itself, it does not offer an opportunity of norm-generating, but it certainly delineates the scope of the application of potential norms, should they be developed by someone else. His argument goes like this:

> "Human beings are not alone in having desires and motives, or in making choices. They share these things with the members of certain other species […] It seems to be peculiarly characteristic of humans, however, that they are able to form what I shall call "second-order desires" or "desires of the second order."

Frankfurt and others, however, are not mistaken in supposing that there *might* be a difference between humans and other animals, or between any two species, for that matter. Take ducks, for example—it is not implausible to think that there is some feature

[97] Frankfurt, *The Problem of Action*, p. 162

specific to them, and them only; something absent in swans, mosquitos and amoebas. Who knows, maybe one day we will identify all the conditions of being a duck, and the list will be so good that it will infallibly sort the universe into ducks and non-ducks. Maybe human beings have an identifiable essence, and thus there is nothing wrong with 'human medium essentialism' argument?

There would be nothing wrong with it if the plausibility of this argument relied on the assumption that human beings are essentially different than other animals. But in fact the validity of this strategy is independent from the question of essences. Sleeping on pillows might well be a distinctly human quality, but that would not mean that we should measure the degree of agency by the number of pillows in our beds. As Carroll puts it, "Of course, even if the doctrine of medium specificity and the sort of essentialism it espouses are false, it still might be the case that cinema has essence."[98]

Let us look at Carroll's argument once again. There is, it seems, no consensus as to what actually constitutes the essence of cinema. But even if there was such agreement, the doctrine of medium essentialism still would not be able to offer relevant guidelines. Is movement essential to movies? Is it the possibility of editing? Particularly compelling mimetic skills? Most likely there will be a rather large set of necessary and sufficient[99] conditions for being classified as a film, including all of the above. Now, if we make the normative assessment dependent on the fulfillment of such heterogeneous 'essence', we

---

[98] Carroll, *Theorizing the Moving Image*, 15.

[99] Elsewhere, Carroll insists that only necessary conditions can be specified (Carroll, *Defining the moving image* in: Carroll and Choi, *Philosophy of Film and Motion Pictures*.

might easily run into trouble. Consider the following problem: which one should be given priority if the mimetic function is disrupted by editorial intervention? And, worse, what if instead of a set of sufficient conditions there is only a list of necessary ones, and within that set some of the conditions are shared with, say, photography?

Similarly, even if we had a list of cognitive processes that are specific to human beings (so far an unlikely empirical achievement, now that elephants have both passed the mirror test and used tools, and there was at least one parrot, who seemed to have a pretty good grasp on semantics)[100], and we decided to treat this list as seriously as medium essentialists, we could conceivably run into trouble. The items on the list could be, for instance, mutually exclusive with respect to the same time-slice, in either logical or physiological sense. It could turn out, for example, that it is *specifically human* to be able to distance oneself from one's desire and evaluate it before it becomes effective, but also to enter, and act under, hypnosis.[101] It could be the case, moreover, that the conjunction of these two conditions forms one sufficient condition for being a human, but each of the conjuncts also belongs to a different set of jointly sufficient conditions (for ferrets and for Moomins, respectively). In that situation the 'essence', even though existent, is useless as a practical recommendation. If we follow the first condition, we are no different than ferrets, if we follow the second, we are just like Moomins; the two

[100] Plotnik, Waal, and Reiss, "Self-recognition in an Asian Elephant." Pepperberg, *Alex & Me*.

[101] I am not sure whether the first feature is true only about humans; the second one is definitely more widespread than that. I chose this as a hypothetical example because these two states are both logically and physiologically exclusive.

combined together are impossible to consistently follow. The sheer possibility of this situation discredits the entire strategy even without any empirical (or commonsensical for that matter) evaluation of the alleged 'essence' postulated by various tortoises.

There is yet another flaw in this strategy, which is best visible through the prism of the Darwinian framework. Let me return to Carroll again: he accurately points out that different artistic media have not been born out of nowhere, but rather developed as extrapolation from another kind of art. Their distinctiveness at a particular moment is more a matter of the course of history of artistic employment. The chaotic practice of art that takes advantage of the existing tools, including available media, and utilizes it for the sake of new manifestations of aesthetic value; this is how 'progress' in art can take place. To come back to our main problem: perhaps we could gain more as moral beings if we spent less time trying to pin down what actions can get a stamp of approval from our metaphysical destiny, and more time finding new paths of being better, even if that meant exploiting the more 'beastly' parts of ourselves. But "…The medium specificity theory maximizes purity instead of excellence,"[102] claims Carroll, and adds a naturalistic, indeed a Darwinian touch to this thought, saying, "…the nature of the medium does not have any determinate directive force concerning the way in which that medium is to be developed." And "It is the use we find for the medium that determined what aspect of the medium deserves out attention. The medium is open to our purposes; the medium does not use us for its own agenda."[103] According to Carroll, art creates its forms non-

[102] Carroll, *Theorizing the Moving Image*. 15

[103] Ibid., 13.

teleologically, and the shapes of these forms are a testimony to the accumulated instances of disorganized, free and local searches for artistic deployment of available elements. Indeed, it is similarly hard to claim normative relevance of a species-specific ability once we assume an evolutionary perspective. Art forms (features of species) are facts; purpose comes from usage—and it is hard to imagine that in the era of Darwinian thinking anyone would believe otherwise. But it is also a fact that we find progress to be valuable, even once its measurements are recognized as relative and its vector revealed to be aimed contingently. From this point of view, freezing a form by endowing it with a normative significance hinders the potential of continuous evolution, whether moral or aesthetic.

Let me stay for a moment longer with the evolutionary perspective. Say we rid the essentialist argument of Aquinas' natural law component, a theoretical tool that *a priori* identifies the range of possible agents. Now we are left with a species that From the point of view of a tortoise, As one species under God, we did not need to worry about identity; as biological beings we must accept the fluidity of evolutionary categorization that is inherent in the very concept of species.[104] That fluidity, if taken seriously, exposes three things. (i) suddenly, essences are exposed as contingent actualities in vast 'design space', (ii) since evolution creates features by blind optimization from the existing conditions and not for the excellence of their prospective function, there is no saying from description alone *how* should a particular feature be used, (iii) in fact, insisting on 'proper' usage of features might hinder further adaptation. In other words, if we want to

[104] For a discussion of the fluidity of the concept of species see Dennett (Dennett, *Darwin's Dangerous Idea*).

talk about ourselves as different from animals, thus putting ourselves in the context of biological comparison, we might have to say that (i) we are not that different, and certainly not necessarily better (at least the standard of comparison must be external to the comparison), (ii) no one can tell us we should use opposing thumbs to play piano and not videogames *unless* they already have an independent standard (same with any cognitive ability we might contingently have, like reasoning about ends) and (iii) if we are allowed to explore a variety of ways of behaving (deliberation, habit, automatic pilot, group hypnosis or explorative imagination), we might find out that our moral lives become easier / better / more fulfilling (in terms, of course, of some independent standard).

The last argument against medium essentialism that Carroll makes is *ad personam*: he claims that he came to believe that in film studies medium essentialism was a defense mechanism the scholars of film developed against the threat of discrediting the discipline. As long as there was something unique about film, film scholarship simply seemed to be more justified in its existence. There is a similar fear in metaethics, I reckon — the special status of human beings at times, and to some, seems like a good justification of advancing moral theory, as if any continuity with the animal world compromised the foundations of the project. The motivational structure behind placing the normative weight on the divide between animals and humans is, however, far more complicated. It does a lot of things for a tortoise. It caters to the metaphysical pride that still remains in force in the secular world, it consoles the ever-lasting post-Ptolemaic sorrow; it cures the existential angst that results from seeing our most cherished abilities

as an evolutionary contingency, bringing back the comfort of purposeful metaphysics, exchanging back the universe for cosmos, in the same time securing some form of conservative political order. But, most of all, it justifies—once again—the search for the roots of morality in theory of action. If we relinquish the distinction between us and the beasts, and we claim that even the most deliberate arm-raising is essentially similar with wagging one's tail, we would have to look for yet another way to justify moral standards, and, frankly, since both, Aquinas' God and Moore's good took a leave of absence, we might be running out of places.

* * *

A Tortoise, then, is a creature that speaks in a language of old-fashioned rationality. Her overt claims are reformed, and yet she relies on the system of semantic preferences that can be explained by a tacit acceptance of a generative metaphorical structure that can be best explained as a GOFAI machine. But the merger between this unmerciful device and the folk-psychological need for a centered and unified agent often ends up in implying some kind of operator to the machine, an executor of sorts, who needs to fend off all the other annoying parts of the psyche. The self-justifying metaphorical GOFAI device helps to satisfy her intuitive need for a steady Archimedean point from which we can then lift normativity. But to do that, we must make it logical for the little homunculus to tend to the device rather than to other parts of the hierarchical mechanism. We should do it, because it is what we do, is one answer; we should do it because this is who we are, is another. Both seem to fail to convince anyone who is not one of Letwin's Romantics, or of the Prussian descent.

Part II

Accommodating Stray Hares:

Towards an Empirically Adequate Image of Agency

Now consider whether knowledge is a thing you can process in that way without having it about you, like a man who has caught some wild birds – pigeons or what not – and keeps them in an aviary he has made for them at home. In a sense, of course we might say he "has" them all in the time inasmuch as he possesses them, mightn't we? (…) But in another sense he "has" none of them (…) (S)o now let us suppose that every mind contains a kind of aviary stocked with birds of every sort, some in flocks apart from the rest, some in small groups, and some solitary, flying in any direction among them all (…)[105]

[*Plato, Theaetatus*]

"You must come here for *some* reason"

"Well, I——" Milo began.

"Come now, if you don't have a reason, you must at least have an explanation, or certainly an excuse," interrupted the gateman.

Milo shook his head.

"Very serious, very serious," the gateman said, shaking his head also. … "Wait a minute, maybe I have an old one you can use"

He took a battered suitcase from the gatehouse and began to rummage busily through it, mumbling to himself: "No… no… this won't do… no… h-m-m-m… ah this is fine," he cried triumphantly, holding up a small medallion on a chain. He dusted it off, and engraved on one side were the words "WHY NOT?"

"That's a good reason for almost anything—a bit used, perhaps, but still quite serviceable"

[*Norton Juster, The Phantom Tollbooth*]

[105] Plato, *Theaetetus*, in: Plato and Cornford, *Plato's Theory of Knowledge; the Theaetetus and the Sophist of Plato*, 197-198a

1. Finding Hares

I have spent a lot of time talking about Tortoises, and how persistent and ubiquitous their species might be. Biologists from Galapagos Islands are undoubtedly surprised at my reading—and, in a sense, rightly so.

For the paradigm that views action-generation mechanism through the lens of a computational metaphor has been bursting at the seams for quite some time now. A lot of thinkers expressed their Hare sympathies—the interest in the actual mechanics of agency, the acceptance of the possibility of a disorderly architecture of the mind, the willingness to give up the possibility of absolute control over one's own moral destiny through the means of impeccable algorithmic processing. Murdoch, whose call against Tortoises prompted the first part of this thesis, argued with the prescriptive demands of deliberation:

> "This too is why … deliberation at the moment of choice often seems ineffectual. We are obscure to ourselves because the world we see already contains our values and we may not be aware of the slow delicate processes of imagination and will which have put those values there."[106]

Her conviction that however well oiled the von Neumann machine might be, it will likely be useless in the real moments of moral significance was picked up by others. Paul Katsafanas, for instance, observes that there is hardly a place in the practical mind where the device can work without interferences of some sort, phenomenology of deliberation notwithstanding:

---

[106] Murdoch *Existentialists and Mystics,* p. 200

> "The agent experiences herself as having a reflective distance from the attitude and asking herself whether there is a reason to act on it; but, all the while, the attitude influences the agent's reflective thought in ways that she does not grasp."[107]

These facts about moral decision-making lead some scholars to the idea that the Tortoise paradigm must be rejected altogether. Amelie Rorty calls the notion that we are presented with extensionally described, discrete courses of actions among which we must rationally ('rationality' understood in the Good Old Fashioned way) decide, "judicialism."

> "Judicialism is – and is acknowledged to be – radically incomplete. It is incomplete as a moral theory because it does not by itself provide a substantive theory of virtue… And … it is incomplete as a psychological theory because it does not exhaust, nor can it serve as the model for, the many functions of thought in forming appropriate actions."[108]

Others tried to develop an alternative account of what I call, in an attempt to use non-committal language, "the stuff that goes on in the head before action happens". Notably, Velleman offered a view on practical reflection that escaped the traps of 'judicialism' by relying on the natural instinct towards intelligibility, best satisfied with generated on-the-go, backround narrative-like self-description. He writes about the reasons to reject a strict method for practical reasoning:

> "I say that practical reasoning is an experimental discipline. The process of figuring out how we can enact intelligible and authentic versions of ourselves cannot be

[107] Katsafanas, *Activity and Passivity in Reflective Agency*, p. 6

[108] Rorty, 280

> boiled down to a syllogism. It cannot be formalized in a calculus of "practitions", means and ends, or desires and beliefs. We reason practically, in the long run by continually trying out clearer, more coherent and yet more ingenuous ways of being and doing; and there is no substitute for trying them out, which is a process of trial and error"[109]

And, elsewhere,

> "(The agent) is not guided by a quantitative balance of reasons, anyway: he is guided rather by the self-understanding that he gains by bringing to consciousness how he thinks and feels about the alternatives"[110]

Patricia Churchland is another example of a thinker who strives for some sort of Hare revolution—but her *coup d'etat* relies less on loosening the rules of the inferential process of practical reflection, and more on campaigning for a naturalistic input in ethics. She sees morality as a matter of negotiation between instincts and institutions, but claims that "…the relation between social urges and the social practices that serve well-being is not simple, and *certainly not syllogistic*" [111] Her Hare temperament demonstrates itself in her holistic approach to the explanation of practical life, one that she borrows from Hume:

> "Hume understood that he needed to have a subtle and sensible account of the complex relationship between moral decisions on the one hand and the dynamic interaction of mental processes -- motivations, thoughts, emotions, memories, and

---

[109] Velleman, *How We Get Along*, 159.

[110] Ibid., 22.

[111] Churchland, *Braintrust*, 8.

> plans -- on the other"[112]

Her commitment to treating the analysis of the characteristics of the species as the starting point for moral reflection is reminiscent of Dewey's concern with what constitutes sensible morality:

> "Moral principles that exalt themselves by degrading human nature are in fact committing suicide. Or else, they involve human nature in unending civil war, and treat it as a hopeless mess of contradictory forces"[113]

It is not perhaps wrong to say, then, that a strong basis for a Hare theory is out there, somewhere between a revival of Dewey, Velleman's attempts to base practical reflection on an instinctual craving for a story and Rorty's insistence on phronetic accounts of virtue. The problem is, I believe, a different one—namely the stiffness of metaphorical imagination in the conceptual space stretched between folk-psychological frameworks and moral philosophy. It is the lack of the guiding metaphor in the space that daily speech and philosophy overlap, the 'commonsensical' subset of the habits of a reflective mind. If there was a model that was strong enough to permeate the very structure of the way we think about ourselves moving in the world, we would be more confident as Hares and more easily convinced as Hares' readers and students. Perhaps the key to reorganizing the power structure in metaethical world is to borrow those models from philosophy of mind that have already formed around available empirical data, and use them as candidate images for the long project of restructuring our action theory, in

---

[112] Churchland, *Braintrust*, 7. Emphasis mine.

[113] Dewey, *Human Nature and Conduct*, pp. xx

the name of the more accurate formulation of moral principles in the future.

For that reason the reminder of this part is devoted to exposition and analysis of such modeling practices. Perhaps, one can hope, Dewey's ingenious conception of habit as a bridge between organism's physiology and citizen's elaborate practice can be rebuilt with the help of recent research in hard sciences, and the symbiotic way in which the habits, though separate, formed a whole, evolutionarily successful individuals can be explained with reference to a more detailed understanding of biological, moral and cultural homeostasis.

2. Modeling Plato's Aviary—Kurzban, Dennett, Fauconnier

I have been arguing that, despite considerable developments in moral philosophy as well as declarations to the contrary, our thinking of action still bears the imprint of the "shopping image". The moral philosophers of the past used to openly construe the human mind in accordance with the principles later described as GOFAI (good old-fashioned artificial intelligence),[114] which is, on a certain reading at least, simply a fancier way of describing the shopping scenario. The pioneers of GOFAI believed that the human mind could be recreated if (1) we developed a sufficiently complex code to designate all objects (or events) and their relations, (2) created a vast base of individual knowledge coded accordingly, (3) used the input device to simulate the stimuli from the environment, (4) used some sort of symbol manipulation system to accommodate the input and generate the output in the form of proper behavioral responses. A GOFAI

[114] The term was coined by John Haugeland in Haugeland, *Artificial Intelligence*.

artificial shopper would be, therefore, no different from Murdoch's shopper: she would have a stable base of beliefs and goals, she would be presented on the input with the assortment of the particular shop she found herself in, and then, using stable rules of symbolic manipulation she would generate the appropriate response in the form of choosing the 'right' product to get. The two shoppers might differ with respect to the ontological status of their conscious mind, which can be relevant to some other aspects of this discussion,[115] but the important thing is that conscious or not, both shoppers were conceived to use the same process to determine the correct action, even if one of them was unaware of its own inner happenings.

The inadequacy of the GOFAI model, I argued, did not exactly prompt a search for a better one, at least insofar as the majority of moral philosophy is concerned; instead, it mostly resulted in incremental adjustments. Parts of the 'knowledge base' were modeled as more difficultly retrievable than others, inputs were limited to reflect the partiality of the mind's awareness of the context, the rules of symbolic manipulation became slightly fuzzier. But the model remained, as a useful reference, a regulative ideal or a blueprint for analysis. As a tool, it became so deeply ingrained in the realities of doing moral philosophy that even small divergences from the GOFAI model (for instance Arpaly's recognition of the messiness of the database) became known as revolutionary

[115] In particular, some dualistically minded theorists might claim that the GOFAI shopper, even though seeming no different than the original one, is not, in fact, conscious or in possession of any mind at all. For the discussion of this problem see Chalmers, *The Conscious Mind* and Dennett's response to Chalmers (Dennett, *The Zombic Hunch* in: Dennett, *Brainchildren*.)

approaches.

But, it seems, there are alternative models available, born on the borders between disciplines such as psychology, philosophy of mind and neuroscience. They appear to be more empirically adequate, more reliably predictive, more effectively explanatory; they just do not seem to inspire metaethical mainstream enough to become visible in the daily life of moral philosophy. Let us look at them in more detail, and, if they deliver, become their advocates, so they can (perhaps) one day be championed by Richard Rorty's 'poets', and enter the deeper structures that govern the ways of thinking.

We can begin, in fact, long time ago. In the famous passage of *Theaetetus*, Plato sketches a metaphorical model of how the human mind works:

> Now consider whether knowledge is a thing you can process in that way without having it about you, like a man who has caught some wild birds – pigeons or what not – and keeps them in an aviary he has made for them at home. In a sense, of course we might say he "has" them all in the time inasmuch as he possesses them, mightn't we? (…) But in another sense he "has" none of them (…) (S)o now let us suppose that every mind contains a kind of aviary stocked with birds of every sort, some in flocks apart from the rest, some in small groups, and some solitary, flying in any direction among them all (…)[116]

Plato's image of the disorderly aviary, where thoughts and memories (birds) are presented as semi-independent agents, engaged in constant movement and ever-changing relationship with other thoughts, turns out to be in striking agreement with the presently

[116] Plato, *Theaetetus*, in: Plato and Cornford, *Plato's Theory of Knowledge; the Theaetetus and the Sophist of Plato*, 197-198a

reliable results of the neuroscientific research.

Robert Kurzban, for instance, has championed a specific perspective on the way we work as acting and reacting agents that is based on a similar intuition: that the mechanism behind our cognitive affairs result from a number of different, not always interdependent processes. This conception of the mind is called *modularity*[117] and has been first put forth by Fodor,[118] but considerably refined afterwards. To explain his

[117] Modularity is often contrasted with the *connectionist* model, where all functions emerge via a unified process of learning from a homogenous structure or with the *computational* model, which is a kind of GOFAI. Some authors believe that a commitment to modularity is mutually exclusive with the idea plasticity of the brain (Ramachandran and Blakeslee, *Phantoms in the Brain*). See Barrett and Kurzban, "Modularity in Cognition." for the discussion on why it is not the case. Modularity does not necessarily presume anatomical basis for each of the modular processes, unless one understands modularity the way Jerry Fodor did (see the next footnote).

[118] Fodor, *The Modularity of Mind*. Fodor's understanding of the modularity hypothesis, however, has been very narrow—he suggested, for instance, that only 'peripheral' systems of the cognitive machinery (like vision or hearing) are modular; he also claimed that the concept of modularity must be connected to nativism, and that modules work necessarily in an automated manner. He himself later criticized his efforts to explain the working of the mind through his view of modularity in Fodor, *The Mind Doesn't Work That Way*. Fodor's problem with extending the notion of modularity was mainly his idea that 'abductive inference', a prominent sort of our cognitive activity concerned with choosing 'the best explanation' for witnessed phenomena, cannot be (he claimed) conceived in modular terms, as it must engage and process the totality of information indiscriminately. See Barrett and Kurzban, "Modularity in Cognition" and Kurzban

model, Kurzban uses a quote from an unpublished draft of Minsky's *Society of the Mind,*[119] where Minsky paints an image that bears some strong resemblance to Plato's aviary metaphor:

> "The mind is a community of "agents". Each has limited powers and can communicate only with certain others. The powers of mind emerge from their interactions for none of the Agents, by itself, has significant intelligence […] Some of them bear useful knowledge, some of them bear strategies for dealing with other agents, some of them carry warnings or encouragements about how the work of others is proceeding. And some of them are concerned with discipline, prohibiting or "censoring" others from thinking forbidden thoughts.

Modularity, as endorsed by Kurzban, not only assumes that mental phenomena arise from the operation of multiple distinct processes rather than a single undifferentiated one; it also claims that the operation of various modules is *all there is*—in other words, that no one is in charge of coordinating the synchronous operations of the modules, no one switches appropriate modules on and off when necessary, no one controls the parallel systems from a meta-level. How do we, then, perform any action at all, if even the simple movement of one's arm arises from the workings of a number of distinct processes, often oblivious of one another? Kurzban answers this question in the following way:

> "[…] it might seem that the dynamic activation and deactivation of modules is

---

and Aktipis, "Modularity and the Social Mind Are Psychologists Too Self-ish? for a refutation of this claim.

[119] Minsky, *Society Of Mind*.

> complicated, influenced by one's age, current state, current context and so on. How is this symphony of modules coordinated? The short answer is that I don't know, and I don't think anyone knows, and that the answer is that there's no one answer, but that, yes, it's all very interesting."[120]

A module, then, is a sub-process of information processing that might, or might not, affect your action or mental state in a given moment.[121] Emotions, for instance, can be understood as a particular interaction between certain modules: "You see a bear and your modules designed for foraging, mating and pretty much everything else get shut off, and your modules for evasion turn on. "Fear" then, is this process, the suite of reactions that lead to some modules gaining priority over others given the current context."[122]

Why would we want to model the mind in a modular way? First, modularity accounts much better for the empirical phenomena than a connectionist view, or any GOFAI type of framework, no matter how complex and self-referential. Second, it offers

---

[120] Kurzban, *Why Everyone (Else) Is a Hypocrite*, 67.

[121] It is, of course, a fairly vague definition. The following quote might help, though. In opposition to Fodor, who tried to specify the necessary conditions for a process to be considered a module, Barrett and Kurzban write: "What it [a modular structure] will look like in a given case—for example, whether or not it will entail automaticity or encapsulation—depends on the details of the mechanism in question. In short, we agree with Pinker, who argued that modules should be defined by the specific operations they perform on the information they receive, rather than by a list of necessary and sufficient features." (Barrett and Kurzban, "Modularity in Cognition," 628.)

[122] Kurzban, *Why Everyone (Else) Is a Hypocrite*, 67., p. 67

explanations of certain common human behaviors that are more efficient and satisfying than any of the opposing models. And third, it is our only way to escape the (already crippled but still persistent) homunculus, who creeps back into all those models of the mind that assume homogeneity of, or strong interconnectedness across, cognitive processing.

Let me start with the first reason—the empirical adequacy of the modular model, and briefly refer to the facts that were traditionally troublesome to non-modular views of the mind. Patients with 'split brains', whose corpus callosum has been partially, or fully severed[123] often manifest a curious symptom: they are unable to name (or report the presence of) an object shown on the left side of their visual field. When presented with two images in separate visual fields, such as a chicken leg on the right and a pile of snow on the left, they report seeing only the chicken leg.[124] Now, when asked to point at a picture that is related to the viewed image, patients pointed at a picture of a chicken with their right hand (correctly pairing it with the chicken leg image), but their left hand always pointed to an image of a snow shovel. They must have seen it *somehow*, it seems. Curiously, when interrogated about pointing at the snow shovel, they would develop a reason wholly independent of the snow scene input (for instance "I saw a claw and I

[123] Most often corpus callosum is cut in the last-ditch attempt to help patients with severe cases of epilepsy.

[124] These experiments were first conducted by Gazzaniga, LeDoux and Wilson; see: LeDoux, Wilson, and Gazzaniga, "A Divided Mind"; Gazzaniga, LeDoux, and Wilson, "Language, Praxis, and the Right Hemisphere"; Gazzaniga, "Brain Mechanisms and Conscious Experience."

picked the chicken, and you have to clean out the chicken shed with a shovel.") The most striking thing about the response of this sort, as Gazzaniga and Le Doux, the lead authors of these experiments, report, was not the fact that it occurred, but the patient's absolute confidence in his[125] explanation. For Gazzaniga, this was a sufficient 'architectural' evidence that human mind is modular: he concluded that speech (in his view the cornerstone to the conscious experience), verbal reasoning and vision were 'running' separately, and communicated only through the thin fibers of the corpus callosum.[126] Both Gazzaniga and Le Doux concluded that corpus callosotomy simply revealed one of the mechanism of the permanent mechanism of rationalizing confabulation that 'whole-brain' patients also exhibit, namely the 'convenient' separation of informational input from the 'inner narrator'.

Kurzban analyzes countless other experiments where people with their brains intact, behave in a fashion similar to those after brain bisection. The blind spot experiments,[127] the rationalizations of the post-hypnotic suggestions[128] or other instances

---

[125] I am using the pronoun 'his' because this particular response was given by a 15 years old male patient, referred to as 'P. S.'

[126] Gazzaniga, "Brain Modularity." Perhaps it is important to say that researchers like Gazzaniga, Le Doux or Ramachandran treat brain dysfunction not as a distinct phenomenon, the sole focus of their studies; rather, they treat it as a window into the workings of the more standard brains.

[127] Where people are presented a stimulus in their blind spot, and respond to it unknowingly, subsequently rationalizing their response (see Ramachandran and Blakeslee, *Phantoms in the Brain*)

of our inability to discern what stimuli have caused our reactions[129] all suggest that information that certainly enters the system might not be accessible to subsets of that system: one module processes it, hides it from another, but its behavior, modified by this information, functions as an input to the other module (such as decision-making module). Kurzban references the research of Nisbett and Wilson, who conclude that there is a massive evidence that "…there may be little or no direct introspective access to higher order cognitive processes"[130]—a claim that supports the modularity thesis.

The second reason to adopt the modular model of the mind lies in its ability to provide satisfying explanations to the phenomena that have otherwise been philosophically troublesome. Consider, for instance, the problem of self-deception and the related problem of *akrasia*. They have been at the center of attention in moral philosophy not just because they might be morally suspect, but also because they were (and still are) hard to fit into the GOFAI models of the mind, even with liberal 'Freudian' adjustments. More often than 'How blameworthy is an akratos?" we pondered the question, "How is the weakness of the will possible". How does modularity help the case here?

The main feature of the modularity view is the postulated lack of a 'module' that can claim absolute informational access: "A key point is that any given specialized

---

[128] For a discussion of this phenomenon, see Daniel M. Wegner, *The Illusion of Conscious Will*.

[129] For a number of experiments of this sort see Ariely, *The Upside of Irrationality*; Ariely, *The (Honest) Truth About Dishonesty*.

[130] Nisbett, Wilson, "Telling More Than We Can Know," p. 231

computational mechanism—any module—might or might not be connected up to any other module."[131] This insight is the origin of the idea of 'informational encapsulation'—the claim that "…your brain can represent mutually inconsistent things at the same time."[132] Not only it *can* represent contradictory things, but it might, in fact, be bound to do so—for evolutionary reasons. Kurzban says, "Not only some modules work better when they have *less information*, some might work better when they have *wrong* information,"[133] and calls the tendency to ignore certain kinds of information in the process of forming specific self-related beliefs the tendency to be 'strategically wrong'. What is 'strategic' about certain ways of being wrong?

Taylor and Brown coined the term 'positive illusions' to describe the kinds of judgment tendencies that habitually overestimate the self-value, the value of one's possessions or one's relatives. Positive illusions are ubiquitous. It is, I venture a wild guess, *probably* impossible that 94% of college professors are 'above average' as instructors, but this is how many of them rate themselves as such;[134] it is equally impossible that a quarter of all SAT takers are in top 1% in terms of their people skills, and yet this is how many of them think that they are.[135] People skills or teaching

---

[131] Kurzban, *Why Everyone (Else) Is a Hypocrite*, 42.

[132] Ibid., 43.

[133] Ibid., 44.

[134] Cross, "Not Can, But Will College Teaching Be Improved?".

[135] This, and other research on this phenomenon begun with the seminal paper by Shelley Taylor and Jonathan Brown (Taylor and Brown, "Illusion and Well-being."). The more recent treatment

effectiveness are, to be fair, somewhat difficult to judge objectively; we can, however, trust that people know their own faces fairly well—after all they see themselves in the mirror almost every day. One of the funniest 'positive illusions' experiments studied the skill of recognizing one's own face. Subjects were shown three or more pictures of themselves, one of them 'raw' (developed without prior Photoshop treatment), all others airbrushed to make them more wrinkle-free and generally more appealing. Most subjects did not identify the 'raw' picture of themselves, pointing to one of the airbrushed versions instead.[136] Somehow, during the exercise, the knowledge of how they looked was switched off, and the optimistic self-image took over.

Phenomena such as positive illusions—the tendency to see ourselves as smarter, prettier, younger[137] and more socially apt than we actually are, or the optimism of terminal cancer patients are far too common to brush them off as cognitive failures. Taylor and Brown hinted at a fairly high correlation between the strength and scope of

---

of this topic can be found in Taylor and Brown, "Positive Illusions and Well-being Revisited."

[136] Taylor and Brown, "Illusion and Well-being."

[137] Taylor and Brown's bold conclusions regarding the advantageousness of positive illusions of were recently confirmed in a study that looked at the 'illusion of youth' in older people. Retirees who harbored an over-exaggerated youthful bias (of more than 15 years) were far more satisfied with their leisure time, had higher self-esteem, felt better health-wise, and were less easily bored than those who felt as old as they were or who, except for perceived health, manifested only a moderate youthful bias (Gana, Alaphilippe, and Bailly, "Positive Illusions and Mental and Physical Health in Later Life.")

positive illusions and one's general well-being.[138] Kurzban takes it one step further and claims that the bundle of modules that take care of our social life can do their job significantly better when we have skewed, or even wrong information about our mating value or survival prospects.[139] After all, our self-confidence might convince someone about our value as mates, and our own steadfast belief in the possibility of survival, even in the face of terminal disease, might safeguard us from being mistreated by those, who would otherwise not have to fear future retaliation. Thus "…some systems might be engineered specifically *not* to get information from (or send information to) other modules."[140]

In this context, Kurzban mentions a text "…dramatically called 'On the very possibility of self deception'," where the author (not mentioned) discusses two subsystems, which he denotes S1 and S2, in the brain of a person. The question the

[138] Similar conclusions were recently drawn by a group of researchers at the university of Belgrade regarding 'positive illusions' about the past. Apparently the strength of the tendency to skew the narrative history in one's favor (by creating false memories or remodeling true ones) is a fairly good predictor of how adjusted and successful a particular individual is (Žeželj et al., "The Impact of Ego-involvement in the Creation of False Childhood Memories.")

[139] This is perhaps the biggest difference between Kurzban's view of the mind and that of Fodor, even though they both have championed the modularity thesis. Fodor believes that the most important function of the mind is the 'fixation of true beliefs' (Fodor, *The Mind Doesn't Work That Way*.) Kurzban believes that the mind want to fixate evolutionarily helpful beliefs, that those just quite often happen to be the true ones—but not as often as one would normally assume.

[140] Kurzban, *Why Everyone (Else) Is a Hypocrite*, 44.

author asks is this: what if S1 believes one thing, but S2 believes another? "This can't possibly be," Kurzban quotes the author, "Because the person cannot, of course, be both S1 and S2." Kurzban finds it to be an amuzing statement: "I love this," he writes, "especially the "of course".[141]

Perhaps, then, our 'social' module is designed to sometimes ignore our 'data collection' modules, or, in sufficiently socially salient contexts simply overrides them, switches them off the way fear of a roaring grizzly bear shuts off our aesthetic appreciation systems. Either possibility explains pretty well (although does not necessarily justify) the structure of some instances of self-deception. Similarly, an *akratos* can be viewed not as a mysterious creature whose beliefs have dubious epistemic status, but as a bundle of operating processes that might well be independent from one another, thus often producing surprising action results. Rather than a problem for theory of action, *akrasia* becomes a fortunate insight into the mechanics of aagency. When we give up the idea of the cognitive center that controls and reviews all reasoning processes, and replace it with a model of parallel processing systems connected by communication channels the width of which is evolutionarily designed and that can open or shut depending on contexts, the explanatory trouble is over. Kurzban frequently references the fascinating work of Breitenburg and his studies on 'vehicles.'[142] Vehicles are very simple electronic devices that are equipped with a few simple sensors (such as light or heat sensor), which are wired to the motion mechanism in a way that makes the vehicle

[141] Ibid., 67.

[142] Braitenberg, *Vehicles*. *Studies in Synthetic Psychology.*

either ‘pursue’ or ‘avoid’ (usually with a set degree of sensitivity) a particular stimulus. Breitenburg’s strikingly simple vehicles exhibit strikingly complex behaviors, especially in situations that make them, so to speak, torn inside. When a vehicle is designed to avoid light and pursue heat, but heat happens to be available only in a well-lit spot, it might, in turns, approach the spot and run away; a simple change in wiring that adjusts the strength of the vehicle’s response might make it remain still half-way towards the source of heat and light, or, if, say, the heat sensor is adjusted to a lower sensitivity, very slowly leave the lit place. The simple system of informational input, such as the heat sensor, together with its connection to the vehicle’s wheels is precisely what Kurzban would call ‘a module’—a separate process, with encapsulated information and independent operations—but functioning alongside the light-avoidance module in a somewhat symbiotic manner. To an external observer, the vehicle’s behavior seems organic and holistic, with clearly identifiable likes and dislikes; the in-principle isolatability of the two modules does not result in any reason to abandon an intentional stance towards the device. An akratos is fairly well explained (but then again, not necessarily justified) as analogous to Breitenburg’s vehicle: pulled in two distinct directions by two separate systems; the net force of their respective persistence determines, in a hardly mysterious fashion, the outcome of her action. The main difference is that an akratos is not made of two, but countless subsystems of this sort, and their interactions with the environing forces and one another are far more complex.

The third reason to adopt the modularity view is its unique ability[143] to generate

[143] An ability that is shares, perhaps, with an old-fashioned determinism, but not with any of the

language that does not invite the sort of metaphorical imagery that became known as the 'Cartesian theatre'. Cartesian theatre happens when the special center in the brain (which could be conceptualized as the functional area of the stably conscious, anatomically, as the 'true' self, the soul or, metaphorically, as the little man at the control panel) is presented the contents of the mental experiences for consideration. Kurzban is weary of the pitfalls of this dualistic view[144] and warns somewhat dramatically, "Without modularity, dualism creeps in."[145]

A knee-jerk reaction of a philosopher to this worry could be to order him to calm down – after all there are reasons to believe that we have successfully ridden philosophy of the homuncular myth. Already James noticed that, "There is no cell or group of cells in the brain of such anatomical or functional preeminence as to appear to be the keystone or center of gravity of the whole system."[146] But Kurzban is not alone in his worry that most psychology and philosophy of the mind still keeps *implying* a homuncular view, even if only as an element of the metaphorical furniture of the particular frameworks. In *Consciousness Explained,* Dennett expresses the same concern:

---

connectionist or GOFAI theories.

[144] Even if the 'special center' were understood anatomically, not in terms of a 'sould' of sorts, it would still invite a kind of dualism: the passive, perceiving 'flesh' of the bodily and mental chaos has to be animated by the active meta-agent.

[145] Kurzban, *Why Everyone (Else) Is a Hypocrite*, 76.

[146] James, *The Principles of Psychology*.

> "Many theorists would insist that they have explicitly rejected such an obviously bad idea. But as we shall see, the persuasive imagery of the Cartesian Theater keeps coming back to haunt us – laypeople and scientists alike – even after its ghostly dualism has been denounced and exorcised. […] The idea of special center in the brain is the most tenacious bad idea bedeviling our attempts to think about consciousness. […] it keeps reasserting itself, in new guises, and for a variety of ostensibly compelling reasons."[147]

How does the homunculus survive in even in what seem to be the most materialistic accounts out there? By locking the control room, throwing the key out, and committing a quiet yet well-publicized suicide—but, in an act of retaliation, leaving the system of corridors leading up to the room intact. The new explanations travel through the old hallways and pass by the locked doors, wistfully glancing at the sealed entrance, realizing that the entire system would make more sense if it reopened. The system of corridors represents what is projected by the structure of the metaphors we use to speak of the affairs in the mind, which has not changed, despite the absence of what has been tying them together. In presenting dilemmas in moral philosophy we imply the existence of the space within which the said dilemma can be surveyed and solved. In speaking of self-deception, we imply that 'on some level' the agent 'does know A,' but chooses to be blinded with B instead. Notice that any kind of talk that strongly problematizes the possibility of inconsistent responses is at the danger of being 'homuncular' in this sense, simply because the very idea of the possibility of unity prompts the thought of a 'unifying' mechanism, a centralized point of view that has all the necessary informational access. This pitfall is elegantly avoided by the modular perspective: "To the extent that

[147] Dennett, *Consciousness Explained*, 1992, 107.

the mind does consist of separate modules, there is just no reason to talk about what one module believes as being more genuine or real than another one," says Kurzban, and adds that "The next time you hear a psychologist try to talk about what someone "really" believes, you should really not believe the psychologist"[148]

At this point we can safely assume that there is at least a strong possibility that the modular view models the mind in a way that is encouragingly empirically adequate, more effectively explanatory than the 'holistic' models and safe from the old ghosts typically blamed on Descartes. However, even if the modular view gives us a fairly smooth explanation of the *mechanics* of conduct, including some instances of morally puzzling conduct such as self-deception and akrasia, we are still without a framework that would allow for the actual insight into the processes of moral thinking. For even if an akratos is like Breitenburg's vehicle, pulled into two different directions by broadly defined modules and ending up following the one that has been evolutionarily favored to take over in this particular type of context, the vehicle herself can justifiably protest that this explanation fails to touch upon her driving experience. A model that would not fail to do so does not have to (and should not) mimic the way the driving *feels* to the agent—an exercise in successful narration about the phenomenological self has little to do with more scientific self-understanding. It makes sense, after all, that evolution could have made us prefer certain things, or do certain things without giving us a knowing access into the actual reasons *why* we prefer or do things in particular fashion, let alone the

[148] Kurzban, *Why Everyone (Else) Is a Hypocrite*, 72.

mechanics behind the thinking or the doing—for evolution 'cares' only that we do these beneficial things, and not about values such as transparency or self-knowledge. What the modular model, in its broad perspective, lacks, is a more detailed view on the mechanics of the self-phenomenology of a (modular) moral agent and her struggle. Even in a strongly materialistic perspective, there must be a place to explain the ways things *seem* to us, when we are trying to drive our modular vehicles.

But how to provide that explanation without giving in to the intuitions that underlie, structure and control *the way the things seem to us*? Certainly the 'shopping' view of the mechanics of thought in action does not manage to do that. As I argued previously, even the modified versions of it still work from the template of a self-transparent, centered, unified and computationally capable mind (that just sometimes, sadly, happens to be a little *less* transparent, *less* unified or *less* computationally capable). I argued that the work on decision-making in social psychology and economics (by researchers such as Gigerenzer[149] or Tversky and Kahnemann[150]) requires us not to piecemeal engineer our model of thinking, but find an entirely new one. The closest to such empirically accurate, phenomenologically plausible, homunculus-free and modularity-friendly model is one that emerges from the development of the concept of

---

[149] Bouissac and Gigerenzer, "Bounded Rationality"; Gigerenzer and McElreath, "Social Intelligence in Games"; Gigerenzer, *Adaptive Thinking*.

[150] Tversky and Kahneman, "Judgment Under Uncertainty"; Tversky and Kahneman, "The Framing of Decisions and the Psychology of Choice." For a summary of their main points see, Kahneman, *Thinking, Fast and Slow*.

mental spaces by Fauconnier,[151] Dennett's analysis of consciousness that results in his 'multiple drafts' idea[152] and his more recent JITSA model that he developed, drawing on Fauconnier's conception, with Hurley and Adams.[153]

Dennett's multiple drafts model of consciousness starts with the intuition that 'Consciousness is gappy and sparse, and doesn't contain half of what people think is there."[154] Consider experiences of being aware of something only post-factum, such as suddenly realizing you have just passed a relative on the street. We can count the chimes of a clock after they stop, without being aware of even hearing it at the time when the sounds were happening. We can recall details of images that have not entered our awareness at the time of viewing. Like cartoon characters, stepping of a cliff as if the road went on, we often keep walking on air until a sudden awareness of an already 'expired' stimulus hits us, and then only we fall. On the 'Cartesian theatre' account, these are pretty easy to explain: our 'consciousness center' was shown perceptions, including the perception of the sounds of the clock or all the details of the viewed images. The operational center decided on the relevance of allowing these particular stimuli further, into the conscious awareness. The homunculus was shown that the road ends at the edge of the cliff, so when further happenings required the revision of the previous screening process, the discarded stimulus was reconsidered by the audience of the Theatre.

---

[151] Fauconnier, *Mental Spaces*.

[152] Dennett, *Consciousness Explained*, 1992.

[153] Hurley, Dennett, and Adams, *Inside Jokes*.

[154] Dennett, *Consciousness Explained*, 1992, 366.

What happens, however, when the missing stimulus was never presented in the theatre in the first place? In strikingly simple experiments with visual stimuli arranged around the 'blind spot' in our vision, we can observe two disconnected parts of a line merging, across the blind spot, into a continuous figure, stretched through an area where it is physically impossible for us to see.[155] Children often approach game booths and grab the joystick to 'play', without even noticing that they have no control over the game at all, because they failed to put the coin in. When adults are put in situations like this, they generate the illusion of causality with equal ease.[156] In the well-known 'color phi' experiment, subjects are shown a blue dot on the left side of the screen and then, after a short period of blank, white screen, a red dot appears on the right side; the subjects are convinced not only that the dot has moved from the left to the right, but also that it has changed the color to red *mid-way*.[157] The question that emerges in studies such as that of the 'color phi' is how exactly consciousness is able to create a perception that includes the *future* information (the change of color from blue to red) *before* it even happens? Are we to say, asks, among others, Goodman,[158] that the brain is clairvoyant, or should we say that the final perceptions are 'constructed"?

Dennett's response to Goodman's question is that he has constructed a false

---

[155] For instructions on doing it yourself, see Ramachandran, *A Brief Tour of Human Consciousness*.

[156] Wegner and Wheatley, "Apparent Mental Causation."

[157] Kolers and von Grünau, "Shape and Color in Apparent Motion."

[158] Goodman, *Ways of World Making*.

dilemma. People like Goodman, claims Dennett, are tacitly committed to the Cartesian idea: they require that we solve this puzzle by postulating some sort of editing mechanism performed for the benefit of the 'controller' or the 'conscious self'—either a *Stalinesque*, or *Orwellian* one. The Stalinesque editing would happen *before* the perception is allowed on the stage: "In the brain's editing room, located before consciousness, there is a delay, a loop of slack like the tape delay used in broadcasts of "live" programs, which gives the censors in the control room a few seconds to bleep out obscenities before broadcasting the signal"[159] The Orwellian mechanism, on the other hand, is a revisionist one: the audience in the theatre sees the perceptions the way they are, but is then promptly offered an alternative version of the recent perceptual history—which, like a nation exhausted with an authoritarian regime, it gladly accepts, forgetting the truth right away.

Dennett notes that that there is a reason to reject both, the Stalinesque and the Orwellian explanation. First, both of them can account for all the data regarding 'revised' perceptions like color phi, not only the existing set of data, but all the future ones we can ever imagine getting, too. And second, speaking of cognition as if it was a process done for the benefit of someone—the audience, the self, the conscious center of cognitive gravity, or consciousness as such—invites the homuncular view back: "What Goodman overlooks is the possibility that the brain doesn't actually have to go to the trouble of "filling in" anything with "construction" – for no one is looking".[160]

---

[159] Dennett, *Consciousness Explained*, 1992, 120.

[160] Ibid., 127.

The alternative is to assume that there are 'multiple drafts' of data interpretation being created by the brain simultaneously; one of them 'wins' and hence modulates subsequent behavior. How does it 'win', however? What is the mechanism of choosing one of the drafts as the guiding one for conduct? Why is one of the drafts 'probed' by consciousness, and the other ones are discarded? The answer Dennett gives is somewhat similar to Kurzban's reply to the question about the synchronization of the various modules: it is hard to say. The feedback from the environment might create a habit of more frequent emergence of a certain kinds of drafts, but generally speaking all this stuff in the brain is a bit of a Pandemonium. When trying to describe how speech is generated, he also imagines multiple drafts of possible reactions being activated and entering some sort of 'fight'—just like the drafts of perception interpretations do:

> In the Pandemonium model, control is usurped rather than delegated, in a process that is largely undesigned and opportunistic; there are multiple sources for the design "decisions" that yield the final utterance, and no strict division is possible between the marching orders of content flowing from within and the volunteered suggestions for implementation posed by the word-demons. What this brand of model suggests is that in order to preserve the creative role of the thought-*expresser* (something that mattered a good deal to Otto [a skeptical character invented by Dennett]), we have to *abandon* the idea that the thought-*thinker* begins with a determinate thought to be expressed. This idea of determinate content also mattered a good deal to Otto, but something has to give.[161]

In other words, our conscious life looks less like this:

"I should probably save this man from the coming bus. But should I? It would be

[161] Ibid., 241.

> morally good, I guess, but what about that meeting I'm going to? Eh, my duty matters more to me than my career, I think. I'll just grab him, there I go… Hey you!"

And more like this:

> "Beeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeep. . . .
> Yabba-dabba-doo-fiddledy-dee-tiddly-pom-fi-fi-fo-fum. . . .
> And so, how about that?, baseball, don't you know, in point of fact, strawberries, happenstance, okay? That's the ticket. Well, then. . . .
> I'm going to knock your teeth down your throat!
> You big meany!
> Ready any good books lately?
> Your feet are too big!"[162]

Dennett's multiple drafts model is, therefore, not only presenting consciousness as less capacious, continuous and orderly than we are used to—it also takes agential control away from it, showing that the quality of being conscious is imparted onto particular brain processes or data largely randomly, in a way that is underdesigned and mostly habit-formed. He goes as far as to suggest that despite its momentous consequences in terms of communicational and learning abilities, consciousness as we understand is not an innate function of the brain, but a *hack* of the brain, "…largely a product of cultural evolution that gets imparted to brains in early training"[163] created through habits that "…would be entrained by frequently saying to a novice, "Tell me what you are doing,"

[162] Ibid., 247.

[163] Ibid., 219.

and "Tell me why you are doing that,"[164] until a novice gets into the habit of addressing the same requests to himself.[165]

We have, therefore, Kurzban's modular view of the mechanisms that make us roll that explains the contradictory and de-centered ways in which we act. Stretched across a subset of these modules we have the sparse and 'gappy' quality of consciousness as described by Dennett's multiple drafts conception. This sort of model gives us a good explanatory access to the way we process information, and does not in any way assume that somewhere inside the homeostatic Pandemonium a von Neumann machine is hidden. But the GOFAI model gave an account of the subset of information processing, with detailed accounts of how it could go. Can we have a similar account for a mind that is not a von Neumann machine, but an unstable aviary?

The JITSA model is an attempt to describe the ways our thinking works in a way that would fit into and specify the underspecified modular framework. It aims at connecting the description of the neural mechanisms behind cognitive processing to the phenomenological experience of thinking, pondering and judging.

What drives most of the Hare ways of modeling the mind is the powerful intuition that our brains seem to have a problem with handling a lot of information at once. How is the brain to do a good job of perception processing or memory search without either lapsing into a combinatorial explosion or failing to represent important information?

---

[164] Ibid., 220.

[165] This particular example only addresses the part of consciousness that concerns the self-narrative of an agent. But Dennett addresses other aspects too—see, Ibid., 220–240.

There must be, a hare thinks, a fairly well-developed automatic way of discarding the potentially irrelevant data. Such automatic ways—among which the most described so far were heuristics and biases—are, of course, bound to once in a while miss something important,[166] but the truth is that, whatever they are, they have been certainly doing a passable job at moving our lives forward.

The JITSA model describes how the brain performs the cognitive tasks it faces without short-circuiting or overheating from the computational process. JIT—just in time—is a term borrowed from software engineering, and refers to program design such that it performs computation only in the moment it needs to be performed; never ahead of time. 'SA' refers to spreading activation across the brain structure. For the authors of the conception, the tool that takes care of balancing the relevance of considered data against the overload is the "…on-demand creation of mental spaces via the process of spreading activation."[167]

---

[166] Some thinkers, like Gigerenzer, still praise the ingenuity of these mechanisms, and claim that heuristics and biases are, given the energy constraints and computational limitations, a much better tool for decision making than Bayesian computation. Hence his insistence on the development of further heuristic tools of decision making. Others, like Kahnemann, find the existence of such mechanisms to be a sad evidence that we are not as rational as we would like to be; the positive program resulting from such attitude is a call for better self-control and less reliance on thin-slicing and other processes of that sort.

[167] Hurley, Dennett, and Adams, *Inside Jokes*, 97.

What is a mental space? The authors, following Fauconnier,[168] describe it as " a region of working memory where activated concepts and percepts are semantically connected into a holistic situational comprehension model"[169] This definition sounds vaguely familiar to anyone acquainted with some of the most important social and cognitive theories of the our time. Gestalt psychology,[170] for instance, suggested that images and situations are perceived holistically, through a process largely controlled by expectations, without the unprompted awareness of details. Erving Goffman's perspective on social interaction described it as the application of appropriate cultural 'scripts' (which can be understood also as semantic wholes awaiting application in long-term memory) in a mutual effort to organize and understand social life and one's role in it.[171] Goffman's script contains the necessary elements of the furniture of a social institution—e.g. the 'restaurant' script will guide us through the comprehension and behavior during a dinner at 'Applebee's'. Goffman's understanding of cognitive 'frames'[172] as the structures that determine the interpretation of events changed sociology forever: it prompted the development of theories such as ethnomethodology,[173] concerned with unearthing and describing the hidden frames and scenarios we harbor and apply in

---

[168] Fauconnier, *Mental Spaces*.

[169] Hurley, Dennett, and Adams, *Inside Jokes*, 97.

[170] Koffka, *Principles of Gestalt Psychology*.

[171] Goffman, *The Presentation of Self in Everyday Life*.

[172] Goffman, *Frame Analysis*.

[173] Garfinkel, *Studies in Ethnomethodology*.

the processing of socially-mediated information. A frame is considered to be a stable psychological device that negotiates perceptions in order to offer certain perspective: it manipulates salience of particular elements to influence subsequent interpretation, judgment or behavior.

But mental spaces are different from 'scripts' and 'frames' in that they are not stable tools, ready to be applied to an initially confusing set of perceptions. They are not "…data structures resident in long-term memory and ready to use when needed" but are, rather, "constructed during comprehension tasks as well as during abstract and creative thought."[174] What is wrong, however, with understanding the framing problem through the prism of Goffmanian scripts?

> "There is an all too common vision of this distinction that we must vigorously oppose here: the idea that long-term memory is a storehouse of sentence-like things (propositions expressed in the "language of thought") that can be *retrieved* and *moved* (or copied) to a special place, *working memory*... […] First of all… the individuation of content into isolated beliefs (billions of them!) in as artifact of our need, in exposition, to draw attention to focal aspects of the information in long-term memory and should not be taken to imply a GOFAI processing model. More important, in this context, is the mistaken image of working memory as the place where things are sent. The antidote to this vision is to remind yourself that we are developing a spreading activation model: working memory is simply that distributed portion of the vast neural network that is currently *working*, awakened, not dormant"[175]

The problem is, then, that scripts are stable and, generally speaking, not

[174] Hurley, Dennett, and Adams, *Inside Jokes*, 97.

[175] Ibid., 107.

underspecified, whereas mental spaces are built on the go. A 'restaurant' script contains all variations of the table arrangement and customer-waiter interactions, and includes the placement of the forks; a mental space begins by a low-level activation of the general restaurant-related content (such as, perhaps, the presence of stuff and places to sit) that is, as one progresses into the dinner date, gradually specified. Each new activation breeds another one, thus the restaurant space will spread to specification differently from the most general activation 'anchor' when in MacDonald's and differently at L'espalier (which is, according to Zagat, the fanciest place in Boston). The fact that mental spaces are build incrementally is supported by the evidence from the studies of garden-path sentences The studies of linguistic processing, especially the analysis of the ways we generate the meaning of garden-path sentences,[176] suggests that a mental space is built predictively (as the stable script theory would suggest) but also *incrementally*—and this is something that cannot be explained by the framework involving stable semantic templates such as frames or scripts. And if we assume that the progress in mental space construction happens because activation is *spread* through the physiological channels of neuronal connections, not because the next fitting beliefs or interpretations are *logically inferred*, this model suddenly becomes a great explanatory device for understanding the ways of our thought. For it always goes in directions that are hard to predict, and yet, when we retrace it, it never seems entirely random. Hurley et al. quote George Santyana

[176] Kamide, Scheepers, and Altmann, "Integration of Syntactic and Semantic Information in Predictive Processing"; Kamide, Altmann, and Haywood, "The Time-course of Prediction in Incremental Sentence Processing."

to give a metaphorical flesh to the JIT spreading of activation:

> "Perceptions do not remain in the mind, as would be suggested by the trite simile of the seal and the wax […] No, perceptions fall into the brain rather as seeds into a furrowed field or even as sparks into a keg of gunpowder. Each image breeds a hundred more, sometimes slowly and subterraneously, sometimes (as when a passionate train is started) with a sudden burst of fancy."[177]

How many mental spaces are there? A simple mind, such as a lower animal's mind, probably generates only one mental space, the one that pertains to the processing of the environmental stimuli here and now—and, in such case, the very concept of a mental 'space' becomes irrelevant. But we are endowed with the capacity to generate multiple mental spaces. The authors say, "When you hear Hamlet tell Ophelia, "Get thee to a nunnery," you can put this into a mental space that you created to contain that story and thereby avoid coming to believe that Hamlet was telling you where to go."[178] Studies seem to indicate that only one mental space can be fully activated at a time, but that "…we may quickly, and with little effort, slip back and forth between them."[179]

Now, in congruence with the multiple drafts model, the entirety of the information in the activated mental space does not have to be 'conscious': if, as the quote above explained, we ditch the obsolete distinction between long-term memory and working memory, we can speak simply of the *levels of activation* of particular brain structures. A solid anchor of the mental space I am generating for the current moment is the fact that I

---

[177] Hurley, Dennett, and Adams, *Inside Jokes*, 97.

[178] Ibid., 98.

[179] Ibid., 97.

am engaged in the process of writing—and that usually involves me sitting on some sort of chair. Up until this moment, however, I was not aware of the chair underneath me. It was not a part of my conscious experience, and it would remain in the unconscious if I did not start looking for an example to mention in the text I am writing. And yet, if the out-of-my-awareness chair has suddenly disappeared (and I would fall), or was unbeknownst to me quietly replaced with a dragon covered in gold scales (and the sudden sensation of movement under my buttocks forced me to look down), I would be surprised (to say the least). The fact that such feeling of surprise would occur, claim the authors, could mean that the 'belief' that the chair is supporting my body has been this entire time activated to some, even if low, degree. And since the development of a mental space is both, incremental *and* eager, a surprise, or shock might occur at a realization that the all-too-keen process went ahead of itself, and activated a belief that is not confirmed later.

Why does it make sense to adopt the JITSA model? Apart from its unparalleled ability to offer an explanation continuous from the physiological aspect to the phenomenology of thinking, it seems, much like the initial modular view, to elucidate some otherwise puzzling things. The baffling problem of 'akratic believers', people who seem to have gathered sufficient evidence to reject some claim and yet still hold on to it. According to Hurley, Dennett and Adams, two contradictory beliefs held in the same time in the long-term memory posit no computational problems for a mind that does not activate them simultaneously. It is even easier, as it is the case with akratic believers, to entertain the disposition to think that x and a disposition to think that y, where y, by all the rules of common sense, implies $\sim x$. I can, therefore, both 'believe' (in terms of an

activational disposition) that my partner is faithful, and also 'believe' that he has been seen naked on top of another, also pretty naked person. How can that happen? First, I need not to activate *x* and *y* at the same time. But the real problem is not the apparent contradiction between the two beliefs, the problem is the lack of habitual associative path that would make my brain prone to activate both of them during the construction of the same inwardly fractalian activation pattern. If "perceptions fall into the brain as seeds into a furrowed field", then in my particular farming habits certain seeds never produce certain kind of sprouts. The authors explain, "We feel epistemic conflict when there is a contradiction between active belief elements in working memory—only when they are brought into the same working memory space, awakened not transported."[180]

Another encouraging feature of the JITSA model is its evolutionary plausibility. Energy conservation is not an evolutionary goal *per se*,[181] but more often than not it is precisely what best serves the replicatory objective—and just-in-time processing is the most thrifty way of tackling computational tasks. One of the objections that might arise in response to this is the problem of anticipation. Evolution must certainly favor foresight: it seems fairly self-evident that we are successful at enduring as a species mainly because we can anticipate the future quite well. JITSA processing might conserve energy, but, since it progresses only on demand, it appears to be a bad anticipatory system. This worry is, however, unwarranted. In JITSA we still produce countless predictions about the future; the thing is that we are not doing it as the shopping models would suggest, by the

[180] Ibid., 102.

[181] Nothing is a 'goal' in evolution, of course; it is a metaphorical way of speaking.

enumeration of all future possibilities followed by the comparison of their likelihood. Instead, "The expectations we have at hand each are the result of current situation-pertinent thought or recollections of other pertinent-at-the-time thoughts each of which are the result of JITSA"[182] The speed with which we grow activation patters is, from the phenomenological point of view, so high, that we are under the illusion of the completeness of the emerging frames,[183] but in fact the only predictions that we end up with are created on the go, through the associative patterns that direct the spreading of the activation. We are lucky in that even though JITSA generates far less predictions that a

[182] Hurley, Dennett, and Adams, *Inside Jokes*, 102.

[183] Think about being told that your friend just bought a ticket to Warsaw. At first, all that you attend to is the fact that he is going to make a trip, perhaps you are not even making any assumptions at all about what kind of ticket he bought. Or maybe you did, but it would be a very low-level one, perhaps an assumption that would activate the possibility of a plane ticket (since one has to get over the Atlantic to get to Poland). If it does turn out to be a plane ticket, it feels as if the frame was strongly present there from the start; the activation of that belief simply occurs anew or becomes higher, and nothing in the mental space has to be rearranged. It feels as if we have known this from the start. But then, what kind of plane is it? Your mental space can equally eagerly accommodate a Boing and an Airbus; neither one disrupts the processing, and fits in, so to speak, like a glove into the on-the-go building enterprise. The speed with which the next steps of specification can be adopted (provided there is no 'shock' involved and our friend did not get a ticket for an experimental use of a teleportation device) poses as a prori completeness of what is, in fact, an incrementally constructed space.

Bayesian model could, it still generally generates predictions that are highly relevant—but then if it did not, it would be discarded much earlier in our evolutionary history.

But since activation spreads in the brain not the way it does in a Von Neumann machine, but more randomly, and it does so with some eager momentum, we are bound to go too far down the wrong activation path once in a while, and assume either too much or too little about the situation. We thought someone was confessing their love to us, when in fact they were talking about our friend. We thought we know who was the murderer before Sherlock Holmes did, and we were wrong. As Europeans, we were convinced that food and drinks are served on all flights; apparently not in the US—and we end up starving in San Francisco well after midnight. Here is, then, another reason to adopt JITSA that comes from evolutionary plausibility. It seems that, as a fallible device, it must have been equipped with some sort of debugging mechanism in order to lead humans into the brave new world of continuous genetic replication. Hurley, Dennett and Adams' book's subtitle is *Using Humor to Reverse-Engineer the Mind*, and their thesis is that mirth we experience when hearing a good joke is an evolutionary reward for noticing a problem with the currently constructed on-the-go mental space.

Consider, for instance, the following joke by Emo Phillips, cited by the authors: "I got into a fight with a very big guy and he said 'I'm going to mop the floor with your face'. I said 'You'll be sorry' and he said 'Oh yeah? Why?' I said 'Well, you won't be able to get into the corners very well.'"[184] The claim is that what causes mirth in this and similar cases is the fact that up to a certain point (the third line of the joke) our JITSA

[184] Hurley, Dennett, and Adams, *Inside Jokes*, 127.

processing already led us quite deep into some particular direction (in this case, of fight-related associations). And yet at the last moment we are shown that the assumptions were inaccurate and the constructed activation pattern will not be able to accommodate the subsequent stimulus—the response is not confrontational, but both a little cowardly and unexpectedly kind. Mirth is a cleverly constructed reaction that rewards most cases of realization that our current mental space with all its future possibilities of further nested and consistent activation has turned to be badly designed from the start.[185] The amusement, it seems, increases in degree proportionally to how covert a particular directional assumption has been.

Hence, the JITSA model is worthy of consideration, for reasons connected to its empirical adequacy (as indicated by the studies on linguistic processing), evolutionary plausibility (energy conservation and general relevance of on-the-go predictions) modeling structure (a seamless continuum from the neurological level to the phenomenological perspective) and its explanatory power (we can finally understand, for instance, why it is so easy to hold contradictory beliefs or what makes a surprise; we can

[185] This is no place for a fuller exposition of this particular theory of humor, but the authors make a convincing case, showing how the 'debugging' of the not-always-reliable JITSA processing can explain cases from first-person humor, when we realize that the glasses we were looking for are actually right there on our nose) through tickling, comedy, absurd humor to the ever-amusing (for infants, that is) peek-a-boo game. There are, of course, cases where such realization is more of a shock than an occasion for joy (someone friendly turns out to be a murderer, for instance), but *formally* the mechanism such discovery is always pleasant.

begin to understand humor and its advantages and the reason why crime stories are such a pleasant thrill—provided they are not happening to us in real life, of course).

We might be, after all, creatures made of modules that are held together by a homeostatic mechanism, who process their data in a disorganized ways and have little access to the causal mechanisms behind out thoughts and behavior. Perhaps we run after birds rather than organize shelves, and perhaps all feelings of agency are, in one way or another, a simple habit of attribution.

3. Conclusion

I have begun this thesis by pointing to a metaphorical structure that has been governing our thinking of action. I argued that this structure can be best explained as an image of a computing machine with linear processing, and that even though hardly ever overtly endorsed, it remained in philosophy of action as a system that governs the tendencies (inferential and semantic) of in particular theories. I called those, who retained this structural metaphor Tortoises, and argued against a number of their strategies. I often mentioned Dewey as a contrast to this way of thinking, as an ultimate Hare, and I expressed a longing for a theory like his, that would tie together the biology and culture, and explained our actions without prejudice, in terms of the process, without a proclivity to satisfy folk-psychological notions, or normative hopes. And then I presented a couple of models of action that promise a similarly comprehensive account, where naturalistic input is allowed in the explanation of moral notions, and empirical adequacy is a relevant element of the model's reliability.

Let us say, than, that we have successfully thrown out the old shopping model and we already have a suitably adequate and detailed model of generating activity. Let us also say that this new model does account for the fact that there are multiple bodily processes happening in parallel, and these differently-minded submodules are not under the command of one mighty module-king, but rather work alongside; their respective operational control determined by their mutual relations in the face of particular kinds of stimuli—very much like in Breitenberg's 'Vehicles'. An agent is perceived, upon this hypothetical model, as a system of fairly separate parts, where a significant subset of them, in a variety of combinations, can bring about those kinds of activities that we usually deem 'actions'. A large subset of this subset consists of modules of essentially the same kind: they are cognitive units of informational processing (described earlier as mental spaces), designed in such a way that despite their vast variety and ability of instantaneous emergence only one of them can be operational at a time — though they can be switched from "on" to "of" with little effort and in no time. And, once again, our model has been built with the deep regard for the fact, even though there is a hierarchy of modules in terms of their direct ability to influence the final resulting activity, there is no one module that can ever claim full control. It could be a model that is openly materialistic and conceives of the conglomerate that is a human being as the end-product of a convoluted evolutionary history; a result of a fairly stable order emerging from the whimsical substratum of chaos. The emergence of behavioral patterns would be viewed as a microcosm of the emergence of the biological agent—as the tactics that survive the test of homeostatic maintenance in the face of particular stimuli. The model would cease

to suggest that the rational mind is the sole, pure agency that can be rescued from within the muddy building matter—'muddiness' characterizing both, the biological structure *and* the system of strategy-building. One could, in the light of the material from part II, definitely conceive of a model like that.

The problem is, of course, that a mechanistic model of this sort does not seem to be the best tool for further ethical theorizing. Even if, as I have been arguing, it is 'better' than other models in metaethics,[186] than it might not exactly come as a metaethical savior, for it seems not to solve any of our normative problems. In fact it seems to do the exact opposite by exposing one fairly widespread Tortoise way of justifying normativity—namely by deriving it from the alleged structure of action—as obsolete. I have discussed examples of this strategy in Part I: Korsgaard,[187] Velleman,[188] and Bratman[189] were my prime illustrations. Candace Vogler, or Sarah Buss are also proponents of such an argument, claiming that understanding practical rationality can only come from

---

[186] In this particular thought here, despite my own views, I do not need to specify a specific way (in terms of empirical adequacy, relevance, reliability, truthfulness, etc.) in which this model might turn out to be 'better' — all I want to express is what happens once it is deemed to be more suitable by anyone for any of these reasons.

[187] Korsgaard, "Self-Constitution in the Ethics of Plato and Kant"; Korsgaard, "Personal Identity and the Unity of Agency"; Korsgaard, *Self-Constitution*.

[188] Velleman, *How We Get Along*; Velleman, "Practical Reflection."

[189] Bratman, "Reflection, Planning, and Temporally Extended Agency"; Bratman, "Two Problems About Human Agency"; Bratman, "Three Forms of Agential Commitment."

familiarity with what actions 'really' are.[190] And, since teleological interpretations are in general neat, frugal and comforting, the day-to-day minds of all of us are to some extent committed to the idea of normativity born out of 'being human'—and doing things 'how they should be done'. And then, as Dennett would say, another mechanistic, materialistic and contingency-emphasizing model comes into the picture and "…spoils the picnic."[191] Why should we allow it? Metaethics have made it clear on numerous occasions that physics is normatively irrelevant[192]; it can once again designate a *persona non grata*, this time in the form of a model that, despite the relevance of its explanatory level to the questions at hand, does not 'move the (normative) action forward' (as script writers say in Hollywood).

My arguments for building a new model of the sort described above have indeed been strangely lacking a decisive mention of possible substantive advantages for normative inquiry. I do, in Part I, consider Dewey's view on a possibility of a successful normative theory—he believes that all responsible guidance must be rooted in knowing the one to be led. There were also others I have brought up, with a similar view on the

---

[190] Buss, "What Practical Reasoning Must Be If We Act for Our Own Reasons"; "Reasonably Vicious — Candace Vogler | Harvard University Press."

[191] Dennett, *DARWIN'S DANGEROUS IDEA*, 10.

[192] The most striking (and most recent) expression of this concern can be found in Thomas Nagel's latest book (Nagel, *Mind and Cosmos: Why the Materialist, Neo-Darwinian Conception of Nature is Almost Certainly False*.)

success conditions of any ethics.[193] And yet the kind of 'knowing' about the mechanics of human activity that the purported model can provide does not seem to be a convenient steppingstone to normative justifications. Emphasizing the mechanics instead of design, and homogeneity of behavior kinds where previously qualitative differences were postulated, leads to, some might think, a moral paralysis of sorts. Problems like responsibility distribution or questions such as moral improvement seem to be in no way aided or relieved by adopting this view. How does, for instance, a more adequate understanding of the leg bone structure help a ballet teacher to teach about ways to improve the beauty of a pose?

My response to such complaint is to admit: yes, there is not much we can do normatively with a model like this, except to treat it as a constraint for our normative fantasies. But such constraints are, I believe, extremely important to have. It is the luck of an ethicist that there are no visible bones to break—and the lack of an anatomy-ignorant ballet teacher that whatever her operative model of the bone structure, it is probably, in the light of her limited range of experience, empirically adequate, since she is not inflicting harm on her students. To overstretch the analogy: if human lives were bodies, and moral theories were like ballet teachers with actual powers to bend the legs and spines of the dancers, there would be quite a lot of severe injuries—unless the anatomy was adequately understood and considered. My suggestion for modeling behavior does

[193] Dewey, *HUMAN NATURE AND CONDUCT*; Velleman, "Practical Reflection"; Murdoch, *Existentialists and Mystics*.

not yield any normative results *in itself*, but it does put constraints on what kind of moral theory is better suited for animals like us.

How exactly? The following comparisons are examples of suggestions that result from adopting the said kind of model. For instance: a theory that tends to distribute moral responsibility for acts between the agent and her social surroundings is, in this light, better than one depicting an agent as the sole target of answerability. Thus Kantian concern with an agent as the source of both good and evil would give way to a structural perspective, where, in a Foucaultian fashion, they emerge from social formations as much as from agential commitments. For example, Claudia Card offers such an account of oppression: she focuses on how wrongdoing lies not just in openly evil acts, but is systematically campaigned for by common neutrality towards oppressive structures through which such acts are, in a large part, bred.[194] Another way in which adoption of a Hare model discriminates between theories pertains to their recommended conscious process of decision-making. Velleman's impromptu imagining of the next move in the context of the narrative-bound 'emotional cadence'—an exercise that is both, creative, instinctive and habitual in nature[195]—beats game-theoretical decision-maker or Audi's deliberator.[196]

Yet another way in which such a model can help us decide between frameworks

---

[194] Card, *The Atrocity Paradigm*; Card, *Confronting Evils*; Card, "Responsibility Ethics, Shared Understandings, and Moral Communities."

[195] Velleman, *How We Get Along*.

[196] Audi, *Practical Reasoning*.

concerns their perspective on the role of institutions in constructing the moral realm. If genuine actions are, on the grounds of the base model, not quite different from 'shmactions' (to use Enoch's word again), the burden of detecting morally relevant 'units' of behavior fall onto the institutional make-up of a given society or group. There is no metaphysics of agency anymore; just like Arthur Danto gutted the essence of art from an artwork and placed it in the surrounding institutional system of 'artworld',[197] a Hare could gladly take away Davidson's 'hope for a principle' and exchange it for institutional regularities of 'agencyworld'. Perhaps in place of a theory of action that infallibly determines, which activity is an action and which one is not, a set of "markers" can be proposed—indicators that a particular activity should be considered as a morally relevant unit. Conceivably, such view of action could look something like this.

Say X is a name of an activity. Of course without metaphysics we immediately run into a problem: how is X even told apart from a stream of ongoing 'doing'? I would say that X, as a term, is essentially fuzzy, and much like W.V. O. Quine's 'Gavagai' must rely on the corrective tools of uniformization that a linguistic group provides in order to sharpen its reference.[198] Now, instead of worrying about the independent properties of X,

---

[197] Danto, "The Artworld."

[198] In Quine's behavioristic view on language acquisition, a new language user can never be fully certain that she correctly identified a particular term's meaning. If a rabbit runs in front of a linguist visiting a pristine tribe, and he hears the word 'Gavagai', he can equally reasonably assume that it means 'rabbit', or 'running rabbit', or 'rabbit's time-slice', or 'rabbit at noon'… some of these hypotheses will be rejected, as they will not lead to desired results in communications,

an action referee must make a decision based on a number of 'action markers'. They could look something like that:

(a) agent feels / reports feeling in control of X

(b) agent entertains a justification of doing X that persists

(c) agent often repeats X

(d) agent reports / does not seem to have been coerced to do X

(e) From the point of view of the moral code used by the 'referee', the activity is either *grossly* wrong or *uncommonly* good, and it has originated, causally speaking, from the body of an agent in question.

There could be more of these markers, I believe; neither of them is necessary, and no particular subset of them is decisive in terms of a decision. A list of this sort is, quite visibly, culturally contingent and can be amended or changed. But these are only *clues*, and they are mostly clues to our language and the institutional make-up of our agencyworld—not to the nature of the acts in question.

Adopting an adequate Hare model might give us a good tool to adjudicate between more and less viable moral theories, on the basis of their fit with the

---

but some of them will be functionally equivalent, thus leaving a language user without any tools for ultimate de-fuzzying of a term's meaning. But linguistic life goes on unfazed, and I see no problem assuming that potential action-units (such as 'slicing cucumbers' or 'making a salad', or 'poisoning a neighbor with a cucumber salad') are not only identified because of previous linguistic and institutional habits, but also are somewhat fuzzy in what they actually mean. The moral life goes on unfazed anyway. (Quine, *Word and Object.*; Quine, *The Roots of Reference*.)

underdesigned devices that evolution has made us to be. The Hare ethics is ethics without a stable ground, where a well-understood creature has to be supplied with suitable goals and rules, and the nature of these rules emerges from previous (also linguistic) inquiry, creativity of the 'law-maker' and continuous testing. The normative component can become, as Hume intended, logically divorced from the descriptive one, and yet it is still designed *for* the creature from the description. If Breitenburg's vehicle is to drive towards the good, we have to learn which of its sensors to adjust, and how—but that does not mean that we can find the direction of 'goodness' engraved on a plaque in its engine.

Can we do this, are we able to self-understand to the level of biology, and still be moral? Or openly non-metaphysical normativity automatically looses its allure, because we do, in fact, need to, 'Repeat old incantations of humanity fables and legends /…/ repeat great words repeat them stubbornly"?[199] I have no idea, but the good news is that we keep graduating from various old fables ever since we wrote them. And if we do outgrow this one, our ethics will become more realistic and effective. Reasonable habits, bolstered by the institutional context and designed for animals like us, modular, 'gappy', governed by association rather than inference, can help us live together a little more happily. Dewey says of wind, "The same air that under certain conditions ruffles the pool or wrecks buildings, under other conditions purifies the blood and [through speech] conveys thought."[200] Perhaps the same water that, as Fortinbras suggests, divides us from

[199] Herbert, *The Envoy of Mr. Cogito*

[200] Dewey, *Human Nature and Conduct*, 20

each other can, under different conditions, become the medium that connects us—if we only accept each other for who we are.

BIBLIOGRAPHY

Ackrill, John L. “Aristotle on Action.” *Mind* 87, no. 348 (October 1, 1978): 595–601. doi:10.2307/2253695.

Anand, Paul, Graham Hunter, and Ron Smith. “Capabilities and Well-Being: Evidence Based on the Sen-Nussbaum Approach to Welfare.” *Social Indicators Research* 74, no. 1 (January 1, 2005): 9–55.

Arganda, Sara, Alfonso Pérez-Escudero, and Gonzalo G. de Polavieja. “A Common Rule for Decision Making in Animal Collectives Across Species.” *Proceedings of the National Academy of Sciences of the United States of America* 109, no. 50 (December 11, 2012): 20508–20513. doi:10.1073/pnas.1210664109.

Ariely, Dan. *The (Honest) Truth About Dishonesty: How We Lie to Everyone---Especially Ourselves*. Harper, 2012.

———. *The Upside of Irrationality: The Unexpected Benefits of Defying Logic*. Reprint. Harper Perennial, 2011.

Arpaly, Nomy. “On Acting Rationally Against One’s Best Judgment.” *Ethics* 110, no. 3 (April 1, 2000): 488–513. doi:10.1086/233321.

Audi, Robert. *Practical Reasoning*. Taylor & Francis Group, 1989.

Augier, Mie, and James G. March. *Models of a Man: Essays in Memory of Herbert A. Simon*. MIT Press, 2004.

Baker, Judith. “Rationality Without Reasons.” *Mind* 117, no. 468 (October 1, 2008): 763–782. doi:10.2307/20532695.

Banks, William P., and Eve A. Isham. “We Infer Rather Than Perceive the Moment We Decided to Act.” *Psychological Science* 20, no. 1 (January 1, 2009): 17–21. doi:10.1111/j.1467-9280.2008.02254.x.

Bannan, John F. “James’s Joke and the Beginnings of the Science of Emotion.” *History of Philosophy Quarterly* 23, no. 1 (January 1, 2006): 59–77.

Barker, John A. “Audi’s Theory of Practical Reasoning.” *Behavior and Philosophy* 19, no. 2 (October 1, 1991): 49–58. doi:10.2307/27759254.

Barrett, H. Clark, and Robert Kurzban. “Modularity in Cognition: Framing the Debate.” *Psychological Review* 113, no. 3 (July 2006): 628–647. doi:10.1037/0033-295X.113.3.628.

Beer, J.S., R.T. Knight, and M. D’Esposito “Controlling the Integration of Emotion and Cognition: The Role of Frontal Cortex in Distinguishing Helpful from Hurtful Emotional Information.” *Psychological Science* 17, no. 5 (May 2006): 448-453.

Bem, D J. “Self-perception: An Alternative Interpretation of Cognitive Dissonance Phenomena.” *Psychological Review* 74, no. 3 (May 1967): 183–200.

Berker, by Selim. “Particular Reasons.” *Ethics* 118, no. 1 (October 1, 2007): 109–139. doi:10.1086/521586.

Bouissac, Paul. “Bounded Rationality: The Adaptive Toolbox. Edited by Gerd Gigerenzer and Reinhard Selten.” *Quarterly Review of Biology* 77, no. 3 (2002): 364.

Braitenberg, Valentino. *Vehicles: Experiments in Synthetic Psychology*. A Bradford Book, 1986.

Bratman, Michael E. “Reflection, Planning, and Temporally Extended Agency.” *Philosophical Review* 109, no. 1 (January 1, 2000): 35–61. doi:10.2307/2693554.

———. “Responsibility and Planning.” *Journal of Ethics* 1, no. 1 (1997): 27–43.

———. “Three Forms of Agential Commitment: Reply to Cullity and Gerrans.” *Proceedings of the Aristotelian Society* 104 (January 1, 2004): 329–337. doi:10.2307/4545421.

———. “Two Problems About Human Agency.” *Proceedings of the Aristotelian Society* 101, no. 1 (June 2001): 309–326.

Brunero, John. “Instrumental Rationality and Carroll’s Tortoise.” *Ethical Theory and Moral Practice* 8, no. 5 (November 1, 2005): 557–569. doi:10.2307/27504377.

Bruun, Henrik, and Richard Langlais. “On the Embodied Nature of Action.” *Acta Sociologica* 46, no. 1 (March 1, 2003): 31–49.

Buss, Sarah. “Autonomous Action: Self-Determination in the Passive Mode.” *Ethics* 122, no. 4 (July 1, 2012): 647–691. doi:10.1086/666328.

Carroll, Lewis. *Through the Looking Glass*. Create Space Independent Publishing Platform, 2010.

Carroll, Noël. *Theorizing the Moving Image*. Cambridge University Press, 1996.

Carroll, Noël. “The Power of Movies.” *Daedalus* 114, no. 4 (October 1, 1985): 79–103 doi:10.2307/20025011.

Carroll, Noël, and Jinhee Choi. *Philosophy of Film and Motion Pictures: An Anthology*. John Wiley & Sons, 2009.

Caspary, William R. “Dewey and Sartre on Ethical Decisions: Dramatic Rehearsal Versus Radical Choice.” *Transactions of the Charles S. Peirce Society* 42, no. 3 (July 1, 2006): 367–393.

Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory*. 1st ed. Oxford University Press, 1997.

Churchland, Patricia S. *Braintrust: What Neuroscience Tells Us About Morality*. Princeton University Press, 2011.

Cox, Damian. “Agent-based Theories of Right Action.” *Ethical Theory and Moral Practice* 9, no. 5 (November 1, 2006): 505–515. doi:10.2307/27504422.

Cross, K. Patricia. “Not Can, But Will College Teaching Be Improved?” *New Directions for Higher Education* (1977). http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ159070.

Cunning, David. “Agency and Consciousness.” *Synthese* 120, no. 2 (January 1, 1999): 271–294. doi:10.2307/20118202.

Dadlez, E. M. *Mirrors to One Another: Emotion and Value in Jane Austen and David Hume*. 1st ed. Wiley-Blackwell, 2009.

Damasio, Antonio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. 1st ed. Mariner Books, 2000.

Dancy, Jonathan. “Defending Particularism.” *Metaphilosophy* 30, no. 1–2 (1999): 25–32. doi:10.1111/1467-9973.00110.

———. *Moral Reasons*. Wiley, 1993.

Davidson, Donald. *Essays on Actions and Events. Underlining*. Oxford University Press, 1980.

Davis, Philip E. “Action‘ and ’Cause of Action.” *Mind* 71, no. 281 (January 1, 1962): 93–95. doi:10.2307/2251736.

Dennett, Daniel C. *Brainchildren: Essays on Designing Minds*. 1st ed. A Bradford Book, 1998.

———. *Consciousness Explained*. 1st ed. Back Bay Books, 1992.

Dewey, John. *Human Nature and Conduct: An Introduction to Social Psychology*. Cosimo Classics, 2007.

Dijksterhuis, Ap, and Loran F. Nordgren. A Theory of Unconscious Thought." *Perspectives on Psychological Science* 1, no. 2 (2006): 95–109.

Döring, Sabine A. "Explaining Action by Emotion." *Philosophical Quarterly* 53, no. 211 (April 1, 2003): 214–230. doi:10.2307/3542865.

Double, Richard. "How to Accept Wegner's Illusion of Conscious Will and Still Defend Moral Responsibility." *Behavior and Philosophy* 32, no. 2 (January 1, 2004): 479–491. doi:10.2307/27759498.

Dworkin, Gerald. "Unprincipled Ethics." *Midwest Studies in Philosophy* 20, no. 1 (1995): 224–239.

Enoch, David. "Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action." *Philosophical Review* 115, no. 2 (April 1, 2006): 169–198. doi:10.2307/20446897.

Fauconnier, Gilles. *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge University Press, 1994.

Fauconnier, Gilles, and Mark Turner. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, 2008.

Feldman, Richard, and Andrei A. Buckareff. "Reasons Explanations and Pure Agency." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 112, no. 2 (January 1, 2003): 135–145. doi:10.2307/4321333.

Fellows, Lesley K., and Martha J. Farah. "The Role of Ventromedial Prefrontal Cortex in Decision Making: Judgment Under Uncertainty or Judgment Per Se?" *Cerebral Cortex* 17, no. 11 (November 1, 2007): 2669–2674. doi:10.1093/cercor/bhl176.

Fesmire, Steven. "Morality as Art: Dewey, Metaphor, and Moral Imagination." *Transactions of the Charles S. Peirce Society* 35, no. 3 (July 1, 1999): 527–550.

———. "Philosophy Disrobed: Lakoff and Johnson's Call for Empirically Responsible Philosophy." *Journal of Speculative Philosophy* 14, no. 4. New Series (January 1, 2000): 300–305.

———. "Dramatic Rehearsal and the Moral Artist: A Deweyan Theory of Moral Understanding." *Transactions of the Charles S. Peirce Society* 31, no. 3 (July 1, 1995): 568–597.

Fodor, Jerry A. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. 1st ed. A Bradford Book, 2001.

———. *The Modularity of Mind: An Essay on Faculty Psychology*. A Bradford Book / MIT Press, 1983.

Foley, Richard. "Deliberate Action." *Philosophical Review* 86, no. 1 (January 1, 1977): 58–69. doi:10.2307/2184162.

Foot, Philippa. *Natural Goodness*. Oxford University Press, 2003.

———. *Virtues & Vices, & Other Essays in Moral Philosophy*. University of California Press, 1978.

Foucault, Michel. *History of Madness*. Edited by Jean Khalfa. Translated by Jonathan Murphy. 1st ed. Routledge, 2006.

———. *Madness and Civilization: A History of Insanity in the Age of Reason.* 1st ed. Vintage, 1988.

Frankfurt, Harry G. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68, no. 1 (January 14, 1971): 5–20. doi:10.2307/2024717.

Gana, K, D Alaphilippe, and N Bailly. "Positive Illusions and Mental and Physical Health in Later Life." *Aging & Mental Health* 8, no. 1 (January 2004): 58–64. doi:10.1081/13607860310001613347.

Garfinkel, Harold. *Studies in Ethnomethodology*. Prentice Hall, 1967.

Garthoff, Jon. "The Embodiment Thesis." *Ethical Theory and Moral Practice* 7, no. 1 (March 1, 2004): 15–29. doi:10.2307/27504293.

Gazzaniga, Michael S. "Brain Mechanisms and Conscious Experience." In *Experimental and Theoretical Studies of Consciousness*. (Ciba Foundation Symposium 174), 1993.

———. "Brain Modularity: Toward a Philosophy of Conscious Experience." In A.J. Marcel and E. Bisiach (eds.) *Consciousness in Contemporary Science*. Oxford University Press, 1988.

———. "Right Hemisphere Language Following Brain Bisection: A 20-year Perspective." *American Psychologist* 38, no. 5 (1983): 525–537. doi:10.1037/0003-066X.38.5.525.

Gazzaniga, Michael S., J. E. LeDoux, and David H. Wilson. "Language, Praxis, and the Right Hemisphere: Clues to Some Mechanisms of Consciousness." *Neurology* 27

(1977): 1144–1147.

Gigerenzer, Gerd. *Adaptive Thinking: Rationality in the Real World*. Oxford University Press, 2002.

Gigerenzer, Gerd, and Richard McElreath. "Social Intelligence in Games: Comment." *Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift Für Die Gesamte Staatswissenschaft* 159, no. 1 (March 1, 2003): 188–194.

Goffman, Erving. *Frame Analysis: An Essay on the Organization of Experience*. Harper & Row, 1974.

———. *The Presentation of Self in Everyday Life*. 1st ed. Anchor, 1959.

Goodman, Nelson. *Ways of World Making*. Hackett Publishing, 1978.

Greene, Joshua D. and Joseph M. Paxton. "Patterns of Neural Activity Associated with Honest and Dishonest Moral Decisions." *Proceedings of the National Academy of Sciences of the United States of America* 106, no. 30 (July 28,2009): 12506–12511.

Haidt, Jonathan. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Reprint. Vintage, 2013.

Haugeland, John. *Artificial Intelligence: The Very Idea*. A Bradford Book, 1989.

Herbert, Zbigniew. *The Collected Poems: 1956-1998*. Reprint. Ecco, 2008.

Hester, Robert, and Hugh Garavan. "Executive Dysfunction in Cocaine Addiction: Evidence for Discordant Frontal, Cingulate, and Cerebellar Activity." *Journal of Neuroscience* 24, no. 49 (December 8, 2004): 11017–11022. doi:10.1523/JNEUROSCI.3321-04.2004.

Higgins, E.T. "How Self-Regulation Creates Distinct Values: The Case of Promotion and Prevention Decision Making." *Journal of Consumer Psychology* 12 (2002): 177–191.

Horgan, Terry, and Mark Timmons. "Morphological Rationalism and the Psychology of Moral Judgment." *Ethical Theory and Moral Practice* 10 (May 2, 2007): 279–295. doi:10.1007/s10677-007-9068-4.

Huffer, Beth. "Actions and Outcomes: Two Aspects of Agency." *Synthese* 157, no. 2 (July 1, 2007): 241–265. doi:10.2307/27653554.

Hume, David. *A Treatise of Human Nature*. Clarendon Press, 1888.

Hurley, Matthew M., Daniel C. Dennett, and Reginald B. Adams Jr. *Inside Jokes: Using Humor to Reverse-Engineer the Mind*. 1st ed. The MIT Press, 2011.

Hutcherson, Cendri A., Hilke Plassmann, James J. Gross, and Antonio Rangel. "Cognitive Regulation During Decision Making Shifts Behavioral Control Between Ventromedial and Dorsolateral Prefrontal Value Systems." *Journal of Neuroscience* 32, no. 39 (September 26, 2012): 13543–13554. doi:10.1523/JNEUROSCI.6387-11.2012.

Huxley, T. "On the Hypothesis That Animals Are Automata, and Its History." *Fortnightly Review* 95 (1874): 555–80.

Isaacs, Tracy Lynn. *Moral Responsibility in Collective Contexts*. Oxford University Press, 2011.

James, William. *Essays in Pragmatism*. Hafner Pub. Co., 1948.

———. *The Principles of Psychology*. New York &: Holt, 1890. http://archive.org/details/theprinciplesofp01jameuoft

Joyce, Richard. "Is Moral Projectivism Empirically Tractable?" *Ethical Theory and Moral Practice* 12 (September 16, 2008): 53–75. doi:10.1007/s10677-008-9127-5.

Juster, Norton. *The Phantom Tollbooth*. Random House, 1989.

Kaag, John. "Chance and Creativity: The Nature of Contingency in Classical American Philosophy." *Transactions of the Charles S. Peirce Society* 44, no. 3 (July 1, 2008): 393–411. doi:10.2307/40321319.

Kahneman, Daniel. *Thinking, Fast and Slow*. 1st ed. Farrar, Straus and Giroux, 2011.

Kamide, Yuki, Gerry T.M. Altmann, and Sarah L Haywood. "The Time-course of Prediction in Incremental Sentence Processing: Evidence from Anticipatory Eye Movements." *Journal of Memory and Language* 49, no. 1 (July 2003): 133–156. doi:10.1016/S0749-596X(03)00023-8.

Kamide, Yuki, Christoph Scheepers, and Gerry T M Altmann. "Integration of Syntactic and Semantic Information in Predictive Processing: Cross-linguistic Evidence from German and English." *Journal of Psycholinguistic Research* 32, no. 1 (January 2003): 37–55.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Cambridge University Press, 1998.

Kearns, Stephen, and Daniel Star. "Reasons: Explanations or Evidence?" *Ethics* 119, no.

1 (October 1, 2008): 31–56.

Kennett, Jeanette, and Cordelia Fine. “Will the Real Moral Judgment Please Stand Up?” *Ethical Theory and Moral Practice* 12 (November 13, 2008): 77–96. doi:10.1007/s10677-008-9136-4.

Kimchi, Eyal Yaacov, and Mark Laubach. “Dynamic Encoding of Action Selection by the Medial Striatum.” *Journal of Neuroscience* 29, no. 10 (March 11, 2009): 3148–3159. doi:10.1523/JNEUROSCI.5206-08.2009.

Koenig, Thomas. “Concepts for Frame Analyses,” 2005. http://www.ccsr.ac.uk/methods/publications/frameanalysis/.

Koffka, K. *Principles of Gestalt Psychology*. Routledge, 1999.

Kolakowski, Leszek. *Metaphysical Horror*. Translated by Agnieszka Kolakowska. 1st ed. University Of Chicago Press, 2001.

Kolers, P A, and M von Grünau. “Shape and Color in Apparent Motion.” *Vision Research* 16, no. 4 (1976): 329–335.

Kolnai, Aurel. “Deliberation Is of Ends.” *Proceedings of the Aristotelian Society* 62 (January 1, 1961): 195–218. doi:10.2307/4544663.

———. *On Disgust*. Edited by Barry Smith and Carolyn Korsmeyer. Open Court, 2004.

Körding, Konrad. “Decision Theory: What ‘Should’ the Nervous System Do?” *Science* 318, no. 5850 (October 26, 2007): 606–610.

Korsgaard, Christine M. “Motivation, Metaphysics, and the Value of the Self: A Reply to Ginsborg, Guyer, and Schneewind.” *Ethics* 109, no. 1 (October 1, 1998): 49–66. doi:10.1086/233873.

Lillehammer, Hallvard. “Analytical Dispositionalism and Practical Reason.” *Ethical Theory and Moral Practice* 2, no. 2 (1999): 117–133.

MacIntyre, Alasdair. *After Virtue: A Study in Moral Theory*, Third Edition. University of Notre Dame Press, 2007.

MacIntyre, Alasdair C. *Dependent Rational Animals: Why Human Beings Need the Virtues (The Paul Carus Lectures)*. Open Court, 2001.

Madden, Rory. “Intention and the Self.” *Proceedings of the Aristotelian Society* 111 (January 1, 2011): 325–351. doi:10.2307/41331555.

Maher, Patrick. “The Irrelevance of Belief to Rational Action.” *Erkenntnis* 24, no. 3

(May 1, 1986): 363–384. doi:10.2307/20012019.

Mather, M, and M K Johnson. “Choice-supportive Source Monitoring: Do Our Decisions Seem Better to Us as We Age?” *Psychology and Aging* 15, no. 4 (December 2000): 596–606.

Melden, A. I. “Action.” *Philosophical Review* 65, no. 4 (October 1, 1956): 523–541. doi:10.2307/2182420.

Mele, Alfred R. “Is Akratic Action Unfree?” *Philosophy and Phenomenological Research* 46, no. 4 (June 1, 1986): 673–679. doi:10.2307/2107677.

Minsky, Marvin. *Society Of Mind.* Simon & Schuster, 1988.

Mitchell, Silas Weir. *Injuries of Nerves and Their Consequences*. Dover Publications, 1965.

Monroe, Kristen. “How Identity and Perspective Constrain Moral Choice.” *International Political Science Review / Revue Internationale de Science Politique* 24, no. 4 (October 1, 2003): 405–425. doi:10.2307/1601630.

Moore, G.E. *Moore Ethics*. Henry Holt And Company, 1907 . http://archive.org/details/mooreethics008304mbp.

Mossel, Benjamin. “Action, Control and Sensations of Acting.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 124, no. 2 (May 1, 2005): 129–180. doi:10.2307/4321600.

Nalini Ambady, Robert Rosenthal. “Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis.” *Psychological Bulletin* 111 (1992): 256–274.

Nietzsche, Friedrich. *Thus Spoke Zarathustra: A Book for Everyone and No One*. Penguin Classics, 1961.

Nussbaum, Martha. *Poetic Justice: The Literary Imagination and Public Life*. Beacon Press, 1997.

———. “Human Functioning and Social Justice: In Defense of Aristotelian Essentialism.” *Political Theory* 20, no. 2 (May 1, 1992): 202–246.

———. *Upheavals of Thought: The Intelligence of Emotions*. Cambridge University Press, 2003.

Paulus, Martin P. “Decision-Making Dysfunctions in Psychiatry-Altered Homeostatic Processing?” *Science* 318, no. 5850 (October 26, 2007): 602–606.

Pepperberg, Irene. *Alex & Me: How a Scientist and a Parrot Discovered a Hidden World of Animal Intelligence--and Formed a Deep Bond in the Process*. Reprint. Harper Perennial, 2009.

Pereboom, Derk. "A Compatibilist Account of the Epistemic Conditions on Rational Deliberation." *Journal of Ethics* 12, no. 3/4 (January 1, 2008): 287–306. doi:10.2307/40345383.

Perry, Thomas D. *Moral Reasoning and Truth: An Essay in Philosophy and Jurisprudence*. 1st Edition. Oxford University Press, 1976.

Pink, Thomas. "Reason and Agency." *Proceedings of the Aristotelian Society* 97 (January 1, 1997): 263–280. doi:10.2307/4545265.

Plassmann, Hilke, John O'Doherty, Baba Shiv, and Antonio Rangel. "Marketing Actions Can Modulate Neural Representations of Experienced Pleasantness." *Proceedings of the National Academy of Sciences of the United States of America* 105, no. 3 (January 22, 2008): 1050–1054. doi:10.1073/pnas.0706929105.

Plato. *Theaetetus*. 1st World Publishing, 2008.

Plato, and Francis Macdonald Cornford. *Plato's Theory of Knowledge; the Theaetetus and the Sophist of Plato*. New York, Liberal Arts Press, 1957. http://archive.org/details/platostheoryofkn00plat.

Plotnik, Joshua M., Frans B. M. de Waal, and Diana Reiss. "Self-recognition in an Asian Elephant." *Proceedings of the National Academy of Sciences of the United States of America* 103, no. 45 (November 7, 2006): 17053–17057. doi:10.1073/pnas.0608062103.

Prinz, Jesse. *The Emotional Construction of Morals*. 1st ed. Oxford University Press, 2009.

Prinz, Jesse J. *Gut Reactions: A Perceptual Theory of Emotion*. Oxford University Press, 2006.

Qin, Lili, Eva M. Pomerantz, and Qian Wang. "Are Gains in Decision-Making Autonomy During Early Adolescence Beneficial for Emotional Functioning? The Case of the United States and China." *Child Development* 80, no. 6 (November 1, 2009): 1705–1721.

Rabinowicz, Wlodek. "Does Practical Deliberation Crowd Out Self-Prediction?" *Erkenntnis* 57, no. 1 (January 1, 2002): 91–122.

Ramachandran, V.S. *A Brief Tour of Human Consciousness: From Impostor Poodles to*

Purple Numbers. Pi Press, 2005.

Ramachandran, V. S., and Sandra Blakeslee. *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. Quill, 1999.

Rayfield, David. "Action." *Noûs* 2, no. 2 (May 1, 1968): 131–145. doi:10.2307/2214701.

Rietveld, Erik. "Situated Normativity: The Normative Aspect of Embodied Cognition in Unreflective Action." *Mind* 117, no. 468 (October 1, 2008): 973–1001. doi:10.2307/20532702.

Robbins, Philip, and Murat Aydede, eds. *The Cambridge Handbook of Situated Cognition*. 1st ed. Cambridge University Press, 2008.

Rogers, Melvin L. "Action and Inquiry in Dewey's Philosophy." *Transactions of the Charles S. Peirce Society* 43, no. 1 (January 1, 2007): 90–115.

Rogers, Robert D., Adrian M. Owen, Hugh C. Middleton, Emma J. Williams, John D. Pickard, Barbara J. Sahakian, and Trevor W. Robbins. "Choosing Between Small, Likely Rewards and Large, Unlikely Rewards Activates Inferior and Orbital Prefrontal Cortex." *Journal of Neuroscience* 19, no. 20 (October 15, 1999): 9029–9038.

Rolls, Edmund T., Fabian Grabenhorst, and Gustavo Deco. "Decision-Making, Errors, and Confidence in the Brain." *Journal of Neurophysiology* 104, no. 5 (November 1, 2010): 2359–2374. doi:10.1152/jn.00571.2010.

Romdenh-Romluc, Komarine. "Agency and Embodied Cognition." *Proceedings of the Aristotelian Society* 111 (January 1, 2011): 79–95. doi:10.2307/41331542.

Rorty, Amélie. "Akratic Believers." *American Philosophical Quarterly* 20, no. 2 (April, 1983): 175–183.

Rorty, Amélie Oksenberg. "Explaining Emotions." *Journal of Philosophy* 75, no. 3 (March 1, 1978): 139–161. doi:10.2307/2025425.

———. *Mind in Action: Essays in the Philosophy of Mind*. Beacon Press, 1991.

———. "Moral Complexity, Conflicted Resonance and Virtue." *Philosophy and Phenomenological Research* 55, no. 4 (December 1, 1995): 949–956. doi:10.2307/2108346.

———. "The Social and Political Sources of Akrasia." *Ethics* 107, no. 4 (July 1, 1997): 644–657.

———. "Three Myths of Moral Theory." In *Mind in Action: Essays in the Philosophy of*

Mind. Beacon Press, 1991.

Rorty, Richard. *Contingency, Irony, and Solidarity*. Cambridge University Press, 1989.

Rosenhan, David. "On Being Sane In Insane Places." *Science* 179, no. 4020 (January 19, 1973): 250–258.

Ross, Steven. "The End of Moral Realism?" *Acta Analytica* 24 (February 5, 2009): 43–61. doi:10.1007/s12136-009-0045-5.

Ross, W.D. *The Right and the Good*. Hackett Publishing Co. Inc., 1988.

Rottschaefer, William A. "Moral Agency and Moral Learning: Transforming Metaethics from a First to a Second Philosophy Enterprise." *Behavior and Philosophy* 37 (January 1, 2009): 195–216. doi:10.2307/41472435.

Russell, L. J. "Ought Implies Can." *Proceedings of the Aristotelian Society* 36 (January 1, 1935): 151–186. doi:10.2307/4544271.

Schapiro, Tamar. "Three Conceptions of Action in Moral Theory." *Noûs* 35, no. 1 (March 1, 2001): 93–117. doi:10.2307/2671947.

Scheler, Max. *Selected Philosophical Essays*. Northwestern University Press, 1992.

Scott, Michael. "Wittgenstein's Philosophy of Action." *Philosophical Quarterly* 46, no. 184 (July 1, 1996): 347–363. doi:10.2307/2956446.

Searle, John R. "How to Derive 'Ought' From 'Is'." *Philosophical Review* 73, no. 1 (January 1, 1964): 43–58. doi:10.2307/2183201.

———. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press, 1983.

Setiya, Kieran. "Explaining Action." *Philosophical Review* 112, no. 3 (July 1, 2003): 339–393. doi:10.2307/3595576.

Sher, George. "Causal Explanation and the Vocabulary of Action." *Mind* 82, no. 325 (January 1, 1973): 22–30. doi:10.2307/2252499.

Shklar, Judith N. *Ordinary Vices*. Harvard University Press, 1984.

Shweder, Richard A. "Everything You Ever Wanted to Know About Cognitive Appraisal Theory Without Being Conscious of It." *Psychological Inquiry* 4, no. 4 (January 1, 1993): 322–326. doi:10.2307/1449649.

Simester, A. P. "Agency." *Law and Philosophy* 15, no. 2 (January 1, 1996): 159–181.

doi:10.2307/3504828.

Simon, Herbert A. *Administrative Behavior*, 4th Edition. Free Press, 2013.

———. "On How to Decide What to Do." *Bell Journal of Economics* 9, no. 2 (October 1, 1978): 494–507. doi:10.2307/3003595.

Simon, Herbert A., Massimo Egidi, Riccardo Viale, and Robin Lapthorn Marris. *Economics, Bounded Rationality and the Cognitive Revolution*. Edward Elgar Publishing, 2008.

Singer, Peter. *Practical Ethics*. 3rd ed. Cambridge University Press, 2011.

Smith, A. D. "Agency and the Essence of Actions." *Philosophical Quarterly* 38, no. 153 (October 1, 1988): 401–421. doi:10.2307/2219706.

Smith, Adam. *The Wealth of Nations*. Simon & Brown, 2012.

Snow, Nancy E. "Habitual Virtuous Actions and Automaticity." *Ethical Theory and Moral Practice* 9, no. 5 (November 1, 2006): 545–561. doi:10.2307/27504425.

Staw, B. "Knee-deep in the Big Muddy: a Study of Escalating Commitment to a Chosen Course of Action." *Organizational Behavior and Human Performance* 16, no. 1 (June 1976): 27–44. doi:10.1016/0030-5073(76)90005-2.

Stevenson, Gordon Park. "Revamping Action Theory." *Behavior and Philosophy* 32, no. 2 (January 1, 2004): 427–451. doi:10.2307/27759495.

Stuss, Donald T., and David Frank Benson. *The Frontal Lobes*. Raven Press, 1986.

Sultana, Mark. *Self-deception and Akrasia: a Comparative Conceptual Analysis*. Gregorian & Biblical Book Shop, 2006.

Tanaka, Saori C., Kazuhiro Shishida, Nicolas Schweighofer, Yasumasa Okamoto, Shigeto Yamawaki, and Kenji Doya. "Serotonin Affects Association of Aversive Outcomes to Past Actions." *Journal of Neuroscience* 29, no. 50 (December 16, 2009): 15669–15674. doi:10.1523/JNEUROSCI.2799-09.2009.

Taylor, S E, and J D Brown. "Illusion and Well-being: a Social Psychological Perspective on Mental Health." *Psychological Bulletin* 103, no. 2 (March 1988): 193–210.

———. "Positive Illusions and Well-being Revisited: Separating Fact from Fiction." *Psychological Bulletin* 116, no. 1 (July 1994): 21–27; discussion 28.

"Telling More Than We Can Know: Verbal Reports on Mental Processes." Accessed April 30, 2013. http://www.batten.virginia.edu/content/facultyresearch/

publications/telling-more-we-can-know-verbal-reports-mental-processes.

Thaler, Richard H., and Prof Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. 1st ed. Yale University Press, 2008.

"The Impact of Ego-involvement in the Creation of False Childhood Memories." Accessed May 4, 2013. http://www.academia.edu/1962806/The_impact_of_egoinvolvement_in_the_creation_of_false_childhood_memories.

Thornton, M. T. "Aristotelian Practical Reason." *Mind* 91, no. 361 (January 1, 1982): 57–76.

Tollefsen, Christopher. "Is a Purely First Person Account of Human Action Defensible?" *Ethical Theory and Moral Practice* 9, no. 4 (August 1, 2006): 441–460. doi:10.2307/27504416.

Tversky, Amos, and Daniel Kahneman. "The Framing of Decisions and the Psychology of Choice." *Science* 211, no. 4481 (January 30, 1981): 453–458. doi:10.1126/science.7455683.

———. "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185, no. 4157 (September 27, 1974): 1124–1131. doi:10.1126/science.185.4157.1124.

Velleman, J. David. *How We Get Along*. Cambridge University Press, 2009.

———. "Practical Reflection." *Philosophical Review* 94, no. 1 (January 1, 1985): 33–61. doi:10.2307/2184714.

Waal, Frans de. *Primates and Philosophers: How Morality Evolved*. Edited by Stephen Macedo and Josiah Ober. Princeton University Press, 2009.

Walker, Nigel. "Freud and Homeostasis." *British Journal for the Philosophy of Science* 7, no. 25 (May 1, 1956): 61–72.

Wallace, R. Jay. "Three Conceptions of Rational Agency." *Ethical Theory and Moral Practice* 2, no. 3 (1999): 217–242.

Wegner, Daniel M. *The Illusion of Conscious Will*. MIT Press, 2002.

Wegner, Daniel M., and T. Wheatley. "Apparent Mental Causation: Sources of the Experience of Will." *American Psychologist* 54 (1999): 480–492.

Wielenberg, by Erik J. "On the Evolutionary Debunking of Morality." *Ethics* 120, no. 3 (April 1, 2010): 441–464. doi:10.1086/652292.

Wiggins, David. “Deliberation and Practical Reason.” *Proceedings of the Aristotelian Society* 76 (1975): 29–51 + viii.

Williston, Byron. “Blaming Agents in Moral Dilemmas.” *Ethical Theory and Moral Practice* 9, no. 5 (November 1, 2006): 563–576. doi:10.2307/27504426.

Wilson, Francis E. “Some Reflections on ‘Action and Reason’.” *Ethics* 83, no. 3 (April 1, 1973): 237–247. doi:10.2307/2380251.

Wittgenstein, Ludwig. *Philosophical Investigations*. New York, Macmillan, 1953.

Wolf, Susan. “Moral Saints.” *Journal of Philosophy* 79, no. 8 (1982): 419–439. doi:10.2307/2026228.

———. “Morality and the View from Here.” *Journal of Ethics* 3, no. 3 (January 1, 1999): 203–223.

Wunderlich, Klaus, Antonio Rangel, and John P. O’Doherty. “Economic Choices Can Be Made Using Only Stimulus Values.” *Proceedings of the National Academy of Sciences of the United States of America* 107, no. 34 (August 24, 2010): 15005–15010. doi:10.1073/pnas.1002258107.

Yolton, John W. “Ascriptions, Descriptions, and Action Sentences.” *Ethics* 67, no. 4 (July 1, 1957): 307–310. doi:10.2307/2379661.

CURRICULUM VITAE

Karolina Lewestam

Leszczyńska 1 / 22

00-339 Warszawa

Karolina.lewestam@gmail.com

Education

- **BA: University of Warsaw, Poland** **1998 - 2004**

College for the Interdisciplinary Studies (MISH). Major: Philosophy. Minor: Social Sciences. GPA: 4.84 (out of 5.0).

- **European College of Liberal Arts, Germany** **2002 – 2003**

GPA: 3.69, completed with distinction

- **PhD: Boston University** **2006 – 2013**
- **Graduate courses taken for credit:**

Philosophy of Law (Prof. D. Lyons), English Empiricism (Prof. W. Hopp), Aristotle's Politics (Prof. D. Roochnik), Philosophy of Human Rights (Prof. A. Biletzki), Political Philosophy (Prof. Simon Keller), Wittgenstein's Investigations (Prof. J. Floyd), Directed Study in Practical Reason and Narrative (Prof. A. Speight), Tort Law (Prof. Simmons)

Philosophical Competence

AOS: Ethics, Philosophy of Action

AOC: Political Philosophy, Philosophy of Mind, Pragmatism, Philosophy of Language, Aesthetics

Teaching Experience

- **Experimental Educational Lab, University of Warsaw 2004 - 2006**

Two years of courses in teaching the humanities including didactics, psychology of learning, methods of interdisciplinary teaching, philosophy of teaching, pedagogy. Degree earned: teaching qualifications.

- **Teaching Assistant at Boston University 2008 - 2011**

Courses: Introduction to Ethics, Introduction to Philosophy, Great Philosophers, Political Philosophy, Philosophy and Film

- **Teaching Assistant at the University of Warsaw 2005**

Courses: Selected Philosophical Problems

- **Lecturer at the Collegium Invisibile Summer Academies 2002 - 2004**

Courses: 'Moral Theory in Practice'; 'Introduction to Language: Philosophy, Linguistics, Psychology.' Taught to talented high school graduates.

- **Teacher at 1st Public High School 'Bednarska', Warsaw 2002**

Courses: Philosophy for 11th grade, 40 hours (covering history of early empirical science, French Enlightenment, Descartes, English empiricism, Kant, German idealism). A review course for 12th grade, 40 hours (a comprehensive preparation for the final "maturity" exam enabling students to enter College)

Selected Publications

Books

• Lewestam, K.: Bajki o Kubusiu. Opcje, Katowice 2009 (A selection of **short stories**)

- Czetwertynska, G., Krawczyk M., Lewestam K.: Pomysl na Szkole z Klasa. Przewodnik po Projektach. Civitas, Warsaw 2008 (**A summary and analysis of the results of a nationwide educational betterment project**)

Articles

- Lewestam, K., Richardson O.: **Why Red Sox Fans are Moral Heroes**. In: "The Red Sox and Philosophy", Open Court Popular Culture and Philosophy Series, 2010
- Casanova i Jego Franciszka, czyli Trzy Razy o Ucieczce Lekkoducha. Kwartalnik Kulturalny Opcje, 02/06 (**Casanova and Francesca, or Three Escapes of a Wanton; an essay on Sandor Marai**)
- Lewestam K., Labenz P., Maslon A.: Project: Agora. In: Biblioteka Prometeusza, Warsaw, 2002 (**Agora: an Educational Project**)
- Lewestam K., Sobala-Zbroszczyk A., Maslon A.: Zawod – Filozof. In Biblioteka Prometeusza, Warsaw 2001 (**Philosopher – a Profession?**)
- 26 Wyprawa Pana Crushera do Innego Wszechswiata. Mishellanea, vol. 2, Warsaw 2001 (**Mr. Crusher's 26th Trip to Another Universe – a mathematical fairly-tale**)
- Lewestam K., Labenz P.: O Quine's Tezie o Niezdeterminowaniu Przekladu Mishellanea, vol. 1, Warsaw 2000 (**On Quine's Thesis of Indeterminacy of Translation**)

Reviews

- Ksztaltujac Sumienie. Przegląd Filozoficzny. Nowa Seria, 15 / 2006 (**Moulding**

your Conscience. A review of Zbigniew Szawarski's Wisdom and the Art of Medicine)

- Philosophy of Logic. Biuletyn Olimpiady Filozoficznej, Warsaw 21/2001 (**review of W. V. O. Quine's book**)
- Theory of Relativity. Biuletyn Olimpiady Filozoficznej, Warsaw, 21/2001 (**review of B. Russell's book**)

Translations

- Manuel Castells, Urban Sustainability in the Information Age, for Kwartalnik Kulturalny 'Opcje', vol. 3, 2010

Academic Awards

- **Boston University Scholarships** (Presidential Scholarship (1st year), Metcalf Fellowship (1st year), Teaching Fellowship (2nd and 5th year), Dissertation Scholarship (4th year))
- **Polish Ministry of Education Scholarship** for outstanding academic performance: 1998 - 1999, 2001 - 2002, and 2003 - 2004)
- **Collegium Invisibile Fellowship and Lifetime Membership** (offered to no more that 20 best Polish students in the humanities yearly): since 2002
- **'Best of the Best'** award for outstanding Polish students: 2002 First prize in the nationwide competition for the **best essay in literature** written for final 'maturity' exam: 1998
- **Polish Academy of Science award** for an essay in philosophy: 1997

Workshops

- **Cinema and Human Rights**: European Inter – University Centre summer university, Venice, 2006
- **International Journalism, Communications, and the Media**: Kokkalis foundation scholarship to Ancient Olympia, 2005
- **Generative Grammar (EGG):** summer school in linguistics, Novi Sad, Serbia, 2002
- **Teacher's Paths through Humanities**: workshops for future educators, Bialobrzegi, Poland, 2001
- **Philosophy of science: Contemporary Controversies in Logic and Metaphysics**. Lvov-Warsaw summer schools, 2001
- **Teaching: a Calling or a Vocation?** Workshops for future educators, Mazury, Poland, 2000

University Service

- Graduate Student Presentation Series (GSPS): organizer in the academic year 2009 - 2010

Work Experience

- **The founding member of Projekt College**, a Liberal Education think-tank: since 2007. Currently involved in launching a teacher-training program in Poland, similar to 'Teach for America'.
- **Editor for Summa Zdarzen, a weekly political show on Polish national TV**: 2004 - 2005

- **Copywriter and presenter for 4FUN TV** (a Polish music TV channel): 2003 – 2004
- Regular freelance work for **G7 Advertising Agency** (strategy department): 2004 – 2006

Languages

• Polish (native) • English (fluent) • German (reading) •

References

Associate Professor **Allen C. Speight**, Boston University, Department of Philosophy (casp8@bu.edu)

Associate Professor **Aaron Garrett**, Boston University, Department of Philosophy (garrett@bu.edu)

Professor Emerita **Amélie Oksenberg Rorty**, Harvard University, Department of Philosophy (Amelie_Rorty@hms.harvard.edu)